LESSLEX: Linking Multilingual Embeddings to SenSe Representations of LEXical Items

Davide Colla
University of Turin - Italy
Computer Science Department
davide.colla@unito.it

Enrico Mensa
University of Turin - Italy
Computer Science Department
enrico.mensa@unito.it

Daniele P. Radicioni
University of Turin - Italy
Computer Science Department
daniele.radicioni@unito.it

We present LESSLEX, a novel multilingual lexical resource. Different from the vast majority of existing approaches, we ground our embeddings on a sense inventory made available from the BabelNet semantic network. In this setting, multilingual access is governed by the mapping of terms onto their underlying sense descriptions, such that all vectors co-exist in the same semantic space. As a result, for each term we have thus the "blended" terminological vector along with those describing all senses associated to that term. LESSLEX has been tested on three tasks relevant to lexical semantics: conceptual similarity, contextual similarity, and semantic text similarity. We experimented over the principal data sets for such tasks in their multilingual and crosslingual variants, improving on or closely approaching state-of-the-art results. We conclude by arguing that LESSLEX vectors may be relevant for practical applications and for research on conceptual and lexical access and competence.

1. Introduction

In the last decade, word embeddings have received growing attention. Thanks to their strength in describing, in a compact and precise way, lexical meaning (paired with a tremendous ease of use), word embeddings conquered a central position in the lexical semantics stage. Thanks to the speed and intensity of their diffusion, the impact of deep architectures and word embeddings has been compared to a tsunami hitting the NLP community and its major conferences (Manning 2015). Word embeddings have been successfully applied to a broad—and still growing—set of diverse application fields,

Submission received: 11 March 2019; revised version received: 5 November 2019; accepted for publication: 29 January 2020.

https://doi.org/10.1162/coli_a_00375

such as computing the similarity between short texts (Kenter and De Rijke 2015), full documents (Kusner et al. 2015), or both (Le and Mikolov 2014). Also, by looking at traditional NLP such as parsing, embeddings proved to be an effective instrument for syntactical parsing—both dependency (Hisamoto, Duh, and Matsumoto 2013; Bansal, Gimpel, and Livescu 2014) and constituency parsing (Andreas and Klein 2014)—and semantic parsing as well (Berant and Liang 2014).

Within this phenomenon, multilingual and crosslingual word embeddings have gained a special status, thanks to the strong and partly unanswered pressure for devising tools and systems to deal with more than one language at a time. Among the main areas where multilingual and crosslingual resources and approaches are solicited, there are of course machine translation (Cho et al. 2014; Luong, Pham, and Manning 2015), crosslingual document categorization (Kočiskỳ, Hermann, and Blunsom 2014; Gouws, Bengio, and Corrado 2015), and sentiment analysis (Tang et al. 2014).

Consistently with the assumption that word semantics is a function of the context (such that words occurring in similar context tend to deliver similar meanings [Harris 1954]), research on word embeddings mostly focused on providing descriptions for terms rather than for word senses, by often disregarding the issue of lexical ambiguity. This fact has historically led (with some exceptions, reviewed hereafter) to a separate growth of research aimed at building word embeddings from that rooted in lexicographic resources (in the tradition of WordNet [Miller 1995] and Babelnet [Navigli and Ponzetto 2010, 2012]) and aimed at developing cognitively plausible approaches to lexical meaning and to the construction of lexical resources. These approaches distinguish between **word meanings** (senses) and **word forms** (terms). The basic unit of meaning is the synset, a set of synonyms that provide (possibly multilingual) lexicalizations to the represented sense, like a lexical dress. Synsets overall describe a semantic network whose nodes are word meanings, linked by semantic relations (such as hypernymy, hyponymy, meronymy, holonymy, etc.). This kind of approach is far in essence from any kind of distributional hypothesis, in that it never happens that a synset conflates two senses.[1] Conversely, the word embeddings for a term provide a synthetic description capturing all senses possibly associated with that term. Word senses have been traditionally used to perform tasks such as word sense disambiguation (WSD) or word sense induction (WSI): individuating which senses occur in a given text may be a precious cue for categorizing documents, to extract meaningful terms, the list of concepts employed along with their mutual relations, etc. The shift of paradigms from lexicographic to distributional approaches has gone hand in hand with the rise of new tasks: besides WSD, semantic similarity (between terms, sentences, paragraphs, whole documents) has emerged as a new, vibrating task in the NLP community: A task perfectly fitting to the geometric descriptions delivered through vectors of real numbers over a continuous, high-dimensional Euclidean space.

1 Adapting to WordNet the definition by Fillmore and Baker about FrameNet, we observe that WordNet is at the "splitting" end of the "splitting" versus "lumping" continuum when it comes to the monosemy/polysemy (Baker and Fellbaum 2009). We presently do not consider the issue of the density of the senses in the sense inventory, which is a relevant issue, with deep consequences on NLP tasks. In fact, whereas fine-grained sense distinctions are necessary for some precise tasks such as machine translation, for other sorts of applications (such as text categorization and information extraction) coarse-grained sense inventories are preferable (Palmer, Babko-Malaya, and Dang 2004). However, the degree of precision needed for any task cannot be algorithmically determined. The issue of filtering the sense inventory received little though significant attention by Navigli (2006), Flekova and Gurevych (2016), Lieto, Mensa, and Radicioni (2016b).

However, despite impressive results obtained by systems using word embeddings, some issues were largely left unexplored, such as (i) the links between representations delivered through word embeddings vs. lexicographic meaning representations; (ii) the cognitive plausibility of the word embeddings (which is different from testing the agreement with conceptual similarity ratings); and (iii) the ways to acquire word embeddings to deliver common-sense usage of language (more on common-sense knowledge later on). In particular, different from lexicographic resources where the minimal addressable unit of meaning is word sense (the synset), with few notable exceptions (such as NASARI [Camacho-Collados, Pilehvar, and Navigli 2015b] and SenseEmbed [Iacobacci, Pilehvar, and Navigli 2015]), word embeddings typically describe terms. This means that different (though close) vectorial descriptions are collected for terms such as *table*, *board*, *desk* for each considered language; whereas in a resource based on senses just one description for the sense of *table* (e.g., intended as "a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs") would suffice. Of course this fact has consequences on the number of vectors involved in multilingual and cross-language applications: One vector per term per language in the case of terminological vectors, one per sense—regardless of the language—otherwise.

One major challenge in the lexical semantics field is, to date, that of dealing with as many as possible languages at the same time (e.g., BabelNet covers 284 different languages),[2] so to enable truly multilingual and crosslingual applications. In this work we propose LESSLEX, a novel set of embeddings containing descriptions for senses rather than for terms. The whole approach stems from the hypothesis that to deal with multilingual applications, and even more in crosslingual ones, systems can benefit from compact, concept-based representations. Additionally, anchoring lexical representations to senses should be beneficial in providing more precise and to some extent more understandable tools for building applications. The evaluation of our vectors seems to support such hypotheses: LESSLEX vectors have been tested in a widely varied experimental setting, providing performances at least on par with state-of-the-art embeddings, and sometimes substantially improving on these.

2. Related Work

Many efforts have been invested in the last decade in multilingual embeddings; a recent and complete compendium is provided by Ruder, Vulić, and Søgaard (2019). In general, acquiring word embeddings amounts to learning some mapping between bilingual resources, so to induce a shared space where words from both languages are represented in a uniform language-independent manner, "such that similar words (regardless of the actual language) have similar representations" (Vulić and Korhonen 2016, page 247). A partially different and possibly complementary approach that may be undertaken is sense-oriented; it is best described as a graph-based approach, and proceeds by exploiting the information available in semantic networks such as WordNet and BabelNet.

2.1 Multilingual Embedding Induction

With regard to the first line of research, in most cases the alignment between two languages is obtained through parallel data, from which as close as possible

2 `https://babelnet.org/stats`.

vectorial descriptions are induced for similar words (see, e.g., the work by Luong, Pham, and Manning [2015]). A related approach consists in trying to obtain translations at the sentence level rather than at the word level, without utilizing word alignments (Chandar et al. 2014); the drawback is, of course, that large parallel corpora are required, which may be a too restrictive constraint on languages for which only scarce resources are available. In some cases (pseudo-bilingual training), Wikipedia has thus been used as a repository of text documents that are circa aligned (Vulić and Moens 2015). Alternatively, dictionaries have been used to overcome the mentioned limitations, by translating the corpus into another language (Duong et al. 2016). Dictionaries have been used as seed lexicons of frequent terms to combine language models acquired separately over different languages (Mikolov, Le, and Sutskever 2013; Faruqui and Dyer 2014). Artetxe, Labaka, and Agirre (2018) propose a method using a dictionary to learn an embedding mapping, which in turn is used to iteratively induce a new dictionary in a self-learning framework by starting from surprisingly small seed dictionaries (a parallel vocabulary of aligned digits), that is used to iteratively align embedding spaces with performances comparable to those of systems based on much richer resources. A different approach consists in the joint training of multilingual models from parallel corpora (Gouws, Bengio, and Corrado 2015; Coulmance et al. 2015).

Also sequence-to-sequence encoder-decoder architectures have been devised, to train systems on parallel corpora with the specific aim of news translation (Hassan et al. 2018). Multilingual embeddings have been devised to learn joint fixed-size sentence representations, possibly scaling up to many languages and large corpora (Schwenk and Douze 2017). Furthermore, pairwise joint embeddings (whose pairs usually involve the English language) have been explored, also for machine translation, based on dual-encoder architectures (Guo et al. 2018).

Conneau et al. (2018) propose a strategy to build bilingual dictionaries with no need for parallel data (MUSE), by aligning monolingual embedding spaces: This method uses monolingual corpora (for source and target language involved in the translation), and trains a discriminator to discriminate between target and aligned source embeddings; the mapping is trained through the adversarial learning framework, which is aimed at acquiring a mapping between the two sets such that translations are close in a shared semantic space. In the second step a synthetic dictionary is extracted from the resulting shared embedding space. The notion of shared semantic space is relevant to our work, which is, however, concerned with conceptual representations. One main difference with our work is that in our setting the sense inventory is available in advance, and senses (accessed through identifiers that can be retrieved by simply querying BabelNet) are part of a semantic network, and independent from any specific training corpus.

For the present work it is important to focus on ConceptNet Numberbatch (CNN hereafter) (Speer and Chin 2016; Speer, Chin, and Havasi 2017). CNN has been built through an ensemble method combining the embeddings produced by GloVe (Pennington, Socher, and Manning 2014) and Word2vec (Mikolov et al. 2013) with the structured knowledge from the semantic networks ConceptNet (Havasi, Speer, and Alonso 2007, Speer and Havasi 2012) and PPDB (Ganitkevitch, Van Durme, and Callison-Burch 2013). CNN builds on ConceptNet, whose nodes are compound words such as "go-to-school." ConceptNet was born with a twofold aim: at expressing *concepts* "which are words and phrases that can be extracted from natural language text," and *assertions* "of the ways that these concepts relate to each other" (Speer and Havasi 2012). Assertions have the form of triples where concept pairs are related by a

set of binary relations:[3] Importantly enough, this knowledge base grasps *common sense*, which is typically hard to acquire by artificial systems. We refer to common sense as a portion of knowledge that is both widely accessible and elementary (Minsky 1975), and reflecting typicality traits encoded as prototypical knowledge (Rosch 1975). This sort of knowledge is about "taking for granted" information, a set of "obvious things people normally know and usually leave unstated" (Cambria et al. 2010, page 15). To the best of our knowledge, no previous system for learning word embeddings has explicitly focused on the acquisition of this sort of knowledge; by contrast, ConceptNet is at the base of other projects concerned with the development of lexical resources (Mensa, Radicioni, and Lieto 2018) and their usage along with formal ontologies (Lieto, Radicioni, and Rho 2015, 2017).

However, ConceptNet is principally a *lexical* resource, and as such it disregards the *conceptual anchoring* issue: If we consider the term *bat*, the *bat* node in ConceptNet mixes all possible senses for the given term, such as the nocturnal mammal, the implement used for hitting the ball, the acronym for "brown adipose tissue," an entity such as the radar-guided glide bomb used by the US Navy in World War II, and so forth.[4] The lack of conceptual anchoring is also a main trait in CNN, as for most word embeddings: Vectors typically flatten all senses, by reflecting their distribution over some corpus approximating human language, or fractions of it.

2.2 Sense Embeddings: Multi-Prototype, Sense-Oriented Embeddings

Some work on word embeddings have dealt with the issue of providing different vectorial descriptions for as many senses associated with a given term. Such approaches stem from the fact that typical word embeddings mostly suffer from the so-called meaning conflation deficiency, which arises from representing all possible meanings of a word as a single vector of word embeddings. The deficiency consists of the "inability to discriminate among different meanings of a word" (Camacho-Collados and Pilehvar 2018, page 743).

In order to account for lexical ambiguity, Reisinger and Mooney (2010) propose representing terms as collections of prototype vectors; the contexts of a term are then partitioned to construct a prototype for the sense in each cluster. In particular, for each word different prototypes are induced, by clustering feature vectors acquired for each sense of the considered word. This approach is definitely relevant to ours for the attempt at building vectors to describe word senses rather than terms. However, one main difference is that the number of sense clusters K in our case is not a parameter (admittedly risking to inject noisy clusters as K grows), but it relies on the sense inventory of BabelNet, which is periodically updated and improved. The language model proposed by Huang et al. (2012) exploits both local and global context that are acquired through a joint training objective. In particular, word representations are computed while learning to discriminate the next word, given a local context composed of a short sequence of words, and a global context composed of the whole document where the word sequence occurs. Then, the collected context representations are clustered, and each occurrence of the word is labeled with its cluster, and used to train the representation for that cluster. The different meaning groups are thus used to learn multi-prototype vectors,

3 The updated list is provided at `https://github.com/commonsense/conceptnet5/wiki/Relations`.
4 `http://conceptnet.io/c/en/bat`.

in the same spirit as in the work by Reisinger and Mooney (2010). Also relevant to our present concerns, the work by Neelakantan et al. (2014) proposes an extension to the Skip-gram model to efficiently learn multiple embeddings per word type: Interestingly enough, this approach obtained state-of-the-art results in the word similarity task. The work carried out by Chen et al. (2015) directly builds on a variant of the Multi-Sense Skip-Gram (MSSG) model by Neelakantan et al. (2014) for context clustering purposes. Namely, the authors propose an approach for learning word embeddings that relies on WordNet glosses composition and context clustering; this model achieved state-of-the-art results in the word similarity task, improving on previous results obtained by Huang et al. (2012) and by Chen, Liu, and Sun (2014).

Another project we need to mention is NASARI. In the same spirit as BabelNet, NASARI puts together two sorts of knowledge: one available in WordNet, handcrafted by human experts, based on synsets and their semantic relations, and one available in Wikipedia, which is the outcome of a large collaborative effort. Pages in Wikipedia are considered as concepts. The algorithm devised to build NASARI consists of two main steps: For each concept, all related Wikipedia pages are collected by exploiting Wikipedia browsing structure and WordNet relations. Then, vectorial descriptions are extracted from the set of related pages. The resource was initially delivered with vectors describing two different semantic spaces: *lexical* (each sense was described through lexical items) and *unified* (each sense was described via synset identifiers). In both cases, vector features are terms/senses that are weighted and sorted based on their semantic proximity to the concept being represented by the current vector (Camacho-Collados, Pilehvar, and Navigli 2015b). In subsequent work NASARI has been extended through the introduction of a distributional description: In NASARI embeddings each item (concept or named entity) is defined through a dense vector over a 300-dimensions space (Pilehvar and Navigli 2015). NASARI vectors have been acquired by starting from the vectors trained over the Google News data set, provided along with the Word2vec toolkit. All the NASARI vectors also share the same semantic space with Word2vec, so that their representations can be used to compute semantic distances between any two such vectors. Thanks to the structure provided by the BabelNet resource, the resulting 2.9*M* embeddings are part of a huge semantic network. Unless differently specified, in the rest of this work we will refer to the embedded version of NASARI, which is structurally more similar to our resource. NASARI includes sense descriptions for nouns, but not for other grammatical categories.

Another resource that is worth mentioning is SENSEEMBED (Iacobacci, Pilehvar, and Navigli 2015); the authors propose here an approach for obtaining continuous representations of individual senses. In order to build sense representations, the authors exploited Babelfy (Moro, Raganato, and Navigli 2014) as a WSD system on the September-2014 dump of the English Wikipedia.[5] Subsequently, the Word2vec toolkit has been used to build vectors for 2.5 millions of unique word senses.

2.3 Contextualized Models

Although not originally concerned with multilingual issues, a mention to works on contextualized embeddings is due, given their large diffusion. Such models are devised to learn dynamic word embeddings representations. Two main strategies can be

5 http://dumps.wikimedia.org/enwiki/.

outlined (Devlin et al. 2019), that apply pre-trained language models to downstream tasks: feature-based and fine-tuning. In the former case, task-specific architectures are used as additional features (like in the case of ELMo [Peters et al. 2018]). Approaches of this sort have been extended to account for sentence (Logeswaran and Lee 2018) and paragraph (Le and Mikolov 2014) embeddings. Peters et al. (2018) extend traditional embeddings by extracting context sensitive features. This kind of model is aimed at grasping complex (such as syntactic and semantic) features associated with word usage, and also to learn how these features vary across linguistic contexts, like in modeling polysemy. ELMo embeddings encode the internal states of a language model based on an LSTM. In the latter case—that is, fine-tuning approaches—minimal task-specific parameters are utilized, and are trained on supervised downstream tasks to tune pre-trained parameters, as in the case of OpenAI GPT (Radford et al. 2019). Unsupervised pre-training approaches are in general known to benefit from nearly unlimited amounts of available data, but approaches exist also showing effective transfer from supervised tasks with large data sets, for example, in sentence representation from NL inference data (Conneau et al. 2017). Specifically, in this work it is shown how universal sentence representations trained using the supervised data of the Stanford Natural Language Inference data sets outperform unsupervised approaches like that by Kiros et al. (2015), and that natural language inference is appropriate for transfer learning to further NLP tasks. BERT vectors (Devlin et al. 2019) rely on other embedding representations, with the notable difference that they model bidirectional context, different from a model such as ELMo, which uses a concatenation of independently trained left-to-right and right-to-left language models.

Until recently, contextualized embeddings of words such as, for example, ELMo and BERT, obtained outstanding performance in monolingual settings, but they seemed to be less suited for multilingual tasks. Aligning contextual embeddings is challenging, because of their dynamic nature. For example, word embeddings tend to be consistent across language variations (Aldarmaki, Mohan, and Diab 2018), whereas multilingual vector spaces have more difficulty in representing individual words (such as, e.g., homographs with unrelated senses and phrasal verbs) because of their different usage distributions. As a result, using such words in the alignment dictionary may undermine the mapping (Aldarmaki and Diab 2019). Another sort of difficulty that may be experienced by contextualized models is represented by cases where a single word of a morphologically complex language corresponds to several words of a morphologically simpler language: In such cases, having a vector for each word might not be appropriate to grasp their meanings across languages (Artetxe and Schwenk 2019, page 3).

However, recent work has been carried out that uses contextual word embeddings for multilingual transfer. The work by Schuster et al. (2019) is reportedly related to MUSE (Conneau et al. 2018): However, different from that approach, aimed at aligning embeddings at the token level, this approach produces alignments for contextual embeddings in such a way that context-independent variants of the original monolingual spaces are built, and their mapping is used to acquire an alignment for context-dependent spaces. More specifically, context-independent embedding anchors are used to learn an alignment that can then be used to map the original spaces with contextual embeddings. With regard to the handling of polysemy, the embeddings obtained through the described approach reflect the multiple senses assumed by the word in different contexts. An alignment based on words in same context, using parallel sentences, is proposed by Aldarmaki and Diab (2019).

3. LessLex Generation

The generation of LessLex relies on two resources: BabelNet and CNN. We briefly describe them for the sake of self-containedness. BabelNet is a wide-coverage multilingual semantic network resulting from the integration of lexicographic and encyclopedic knowledge from WordNet and Wikipedia. Word senses are represented as *synsets*, which are uniquely identified by Babel Synset identifiers (e.g., `bn:03739345n`). Each synset is enriched by further information about that sense, such as its possible lexicalizations in a variety of languages, its gloss (a brief description), and its Wikipedia Page Title. Moreover, it is possible to query BabelNet to retrieve all the meanings (synsets) for a given term. Although the construction of BabelNet is by design essential to our approach, in principle we could plug in different sets of word embeddings. We chose CNN word embeddings as our starting point for a number of reasons, namely: its vectors are to date highly accurate; all such vectors are mapped onto a single shared multilingual semantic space spanning over 78 different languages; it ensures reasonable coverage for general purposes use (Speer and Lowry-Duda 2017); it allows dealing in a uniform way with multiword expressions, compound words (Havasi, Speer, and Alonso 2007), and even flexed forms; and it is released under the permissive MIT License.

The algorithm for the generation of LessLex is based on an intuitive idea: to exploit multilingual terminological representations in order to build precise and punctual conceptual representations. Without loss of generality, we introduce our methodology by referring to nominal senses, although the whole procedure also applies to verb and adjectival senses, so that in the following we will switch between sense and concept as appropriate. Each concept in LessLex is represented by a vector generated by averaging a set of CNN vectors. Given the concept c, we retrieve it in BabelNet to obtain the sets $\{\mathcal{T}^{l_1}(c), \ldots, \mathcal{T}^{l_n}(c)\}$ where each $\mathcal{T}^{l}(c)$ is the set of lexicalizations in the language l for c.[6] We then try to extract further terms from the concepts' English gloss and English Wikipedia Page Title (WT from now on). The final result is the set $\mathcal{T}^{+}(c)$ that merges all the multilingual terms in each $\mathcal{T}^{l}(c)$ plus the terms extracted from the English gloss and WT. In $\mathcal{T}^{+}(c)$ we retain only those terms that can be actually found in CNN, so that the LessLex vector $\vec{c}$ can be finally computed by averaging all the CNN vectors associated to the terms in $\mathcal{T}^{+}(c)$.

3.1 Selecting the Sense Inventory: Seed Terms

Because the generation algorithm creates a representation for conceptual elements (be they nominal, verbal, or adjectival senses), it is required to define which concepts will be hosted in the final resource. For this purpose we define a set of terms that we call *seed terms*. Seed terms are taken from different languages and different POS (nouns, verbs, and adjectives are presently considered), and their meanings (retrieved via BabelNet) constitute the set of senses described by LessLex vectors. Because of the polysemy of language and because the seed terms are multilingual, different seed terms can retrieve the same meaning. Seed terms do not affect the generation of a vector, but they rather determine the coverage of LessLex, since they are used to acquire the set of concepts

6 We presently consider all the languages that are adopted during the evaluation: English (eng), French (fra), German (deu), Italian (ita), Farsi (fas), Spanish (spa), Portuguese (por), Basque (eus), and Russian (rus).

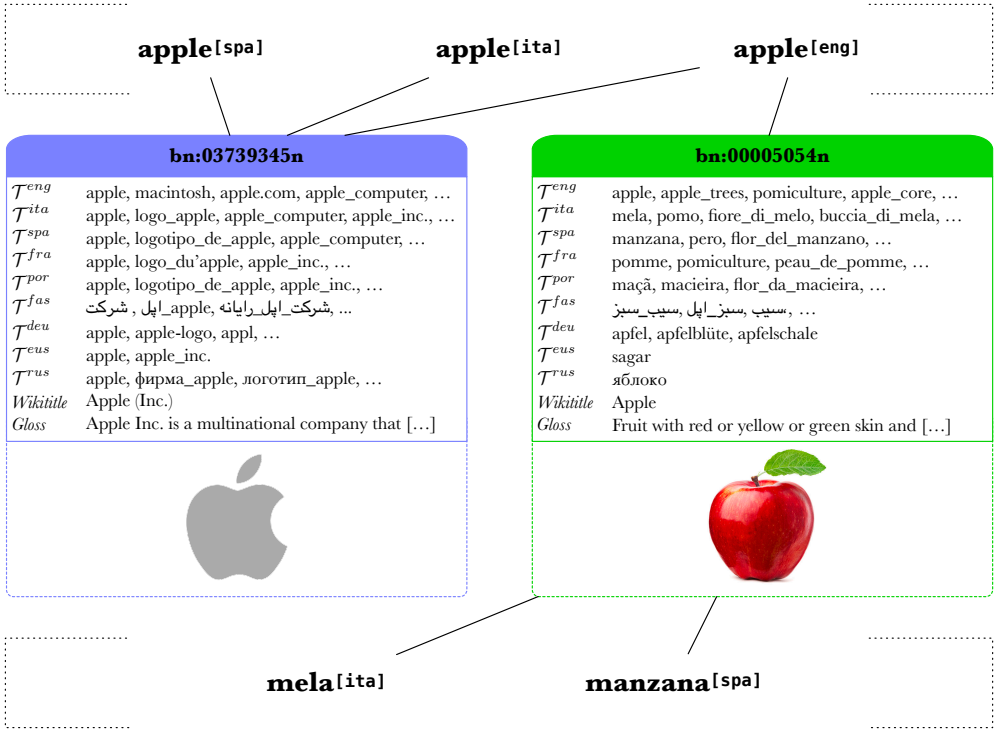

Figure 1
Retrieval of two senses for five seed terms in three different languages.

that will be part of the final resource. Figure 1 illustrates this process for a few seed terms in English, Spanish, and Italian. These terms provide two senses in total: `bn:03739345n` – *Apple (Inc.)* and `bn:00005054n` – *Apple (fruit)*. The first one is the meaning for $apple^{spa}$, $apple^{ita}$, and $apple^{eng}$, and the second one is a meaning for $manzana^{spa}$, $mela^{ita}$, and, again, $apple^{eng}$. Each synset contains all the lexicalizations in all languages, together with the English gloss and the WT. This information will be exploited for building $\mathcal{T}^{+}(c_{bn:03739345n})$ and $\mathcal{T}^{+}(c_{bn:00005054n})$ during the generation process.

3.2 Extending the Set of Terms

As anticipated, we not only rely on the lexicalizations of a concept to build its $\mathcal{T}^{+}$, but we also try to include further specific words, parsed from its English gloss and WT. The motivation behind this extension is the fact that we want to prevent $\mathcal{T}^{+}$ from containing only one element: In such a case, the vector for the considered sense would coincide with that of the more general term, possibly conflating different senses. In other words, enriching $\mathcal{T}^{+}$ with further terms is necessary to reshape vectors that have only one associated term as lexicalization. For instance, starting from the term $sunset^{eng}$ we encounter the sense `bn:08410678n` (representing the city of Sunset, Texas). This sense is provided with the following lexicalizations:

$$\mathcal{T}^{eng} = \{sunset^{eng}\}; \quad \mathcal{T}^{spa} = \{sunset^{spa}\}; \quad \mathcal{T}^{fra} = \{sunset^{fra}\}.$$

However, out of these three terms only *sunset*^{eng} actually appears in CNN, giving us a final singleton $\mathcal{T}^+ = \{sunset^{eng}\}$. At this point no average can be performed, and the final vector in LESSLEX for this concept would be identical to the vector of *sunset*^{eng} in CNN. Instead, if we take into consideration the gloss ''*Township in Starr County, Texas*,'' we can extract *township*^{eng} and append it in $\mathcal{T}^+$, thus obtaining a richer vector for this specific sense of *sunset*. In the following sections we describe the two strategies that we developed in order to extract terms from WTs and glosses. The extension strategies are applied for every concept, but in any case, if the final $\mathcal{T}^+$ contains a single term ($|\mathcal{T}^+| = 1$), then we discard the sense and we do not include its vector in LESSLEX.

3.2.1 Extension Via Wikipedia Page Title. The extension via WT only applies to nouns, because senses for different POSs are not present in Wikipedia. In detail, if the concept has a Wikipedia Page attached and if the WT provides a disambiguation or specification (e.g., *Chips (company)* or *Magma, Arizona*) we extract the relevant component (by exploiting commas and parentheses of the Wikipedia naming convention) and search for it in CNN. If the whole string cannot be found, we repeat this process by removing the leftmost word of the string until we find a match. In so doing, we search for the maximal substring of the WT that has a description in CNN. This allows us to obtain the most specific and yet defined term in CNN. For instance, for the WT *Bat (guided bomb)* we may not have a match in CNN for *guided bomb*, but we can at least add *bomb* to the set of terms in $\mathcal{T}^+$.

3.2.2 Extension Via Gloss. Glosses often contain precious pieces of information that can be helpful in the augmentation of the terms associated with a concept. We parse the gloss and extract its components. By construction, descriptions provided in BabelNet glosses can originate from either WordNet or Wikipedia (Navigli and Ponzetto 2012). In the first case we have (often elliptical) sentences, such as (`bn:00028247n` – *door*) "a swinging or sliding barrier that will close the entrance to a room or building or vehicle." On the other side, Wikipedia typically provides a plain description like "A door is a panel that makes an opening in a building, room or vehicle." Thanks to the regularity of these languages, with few regular expressions on POS patterns[7] we are able to collect enough information to enrich $\mathcal{T}^+$. We devised several rules according to each sense POS; the complete list is reported in Table 1. As an example, from the following glosses we extract the terms in bold (the matching rule is shown in square brackets):

- [Noun-2] `bn:00012741n` (*Branch*) *A* ***stream*** *or* ***river*** *connected to a larger one.*
- [Noun-3] `bn:00079944n` (*Winner*) *The* ***contestant*** *who wins the contest.*
- [Noun-1] `bn:01276497n` (*Plane (river)*) *The Plane is a* ***river*** *in Brandenburg, Germany, left tributary of the Havel.*
- [Verb-2] `bn:00094850v` (*Tee*) ***Connect*** *with a tee.*
- [Verb-3] `bn:00084198v` (*Build*) ***Make*** *by* ***combining*** *materials and parts.*

7 We adopted the Penn Treebank POS set: `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`.

Table 1
List of the extraction rules in a regex style, describing some POS patterns. If a gloss or a portion of a gloss matches the left part of the rule, then the elements in the right part are extracted. Extracted elements are underlined.

Nouns		
1. to be NN+	⟶	NN+
2. NN1 CC NN2	⟶	NN1,NN2
3. DT ∗ NN+	⟶	NN+
Verbs		
1. to be VB	⟶	VB
2. Sentence starts with a VB	⟶	VB
3. VB1 ((CC | ,) VB2)+	⟶	VB1, VB2+
Adjectives		
1. Sentence is exactly JJ	⟶	JJ
2. *not* JJ	⟶	(JJ is dropped)
3. (*relate*|*relating*|*related*) *to* ∗ NN	⟶	NN
4. JJ1 CC JJ2	⟶	JJ1,JJ2
5. JJ1, JJ2 *or* JJ3	⟶	JJ1, JJ2, JJ3

- [Adjective-3] `bn:00106822a` (*Modern*) *Relating to a recently developed* ***fashion*** *or style*.
- [Adjective-4] `bn:00103672a` (*Good*) *Having* ***desirable*** *or* ***positive*** *qualities especially those suitable for a thing specified*.

In Figure 2 we provide an example of the generation process for three concepts, provided by the seed terms $gate^{eng}$ and $gate^{ita}$. For the sake of simplicity, we only show the details regarding two languages (English and Italian). Step (1) shows the input terms. In step (2) we retrieve three meanings for $gate^{eng}$ and one for $gate^{ita}$, which has already been fetched because it is also a meaning for $gate^{eng}$. For each concept we collect the set of lexicalizations in all considered languages, plus the extensions extracted from WT and gloss. We then merge all such terms in $\mathcal{T}^+$, by retaining only those that can be actually found in CNN. Once the $\mathcal{T}^+$ sets are computed, we access CNN to retrieve the required vectors for each set (3) and then we average them, finally obtaining the vectors for the concepts at hand (4).

3.3 LESSLEX Features

We now describe the main features of LESSLEX, together with the algorithm to compute conceptual similarity on this resource. The final space in which LESSLEX vectors reside is an extension of the CNN multilingual semantic space. Each original CNN vector coexists with the set of vectors that represent its underlying meanings. This peculiar feature allows us to compute the distance between a term and each of its corresponding senses, and such distance is helpful to determine, given a pair of terms, in which sense they are intended. For example, in assessing the similarity of two terms such as "glass" and "eye," most probably the recalled senses would differ from those recalled for the pairs "glass" and "window," and "glass," "wine."

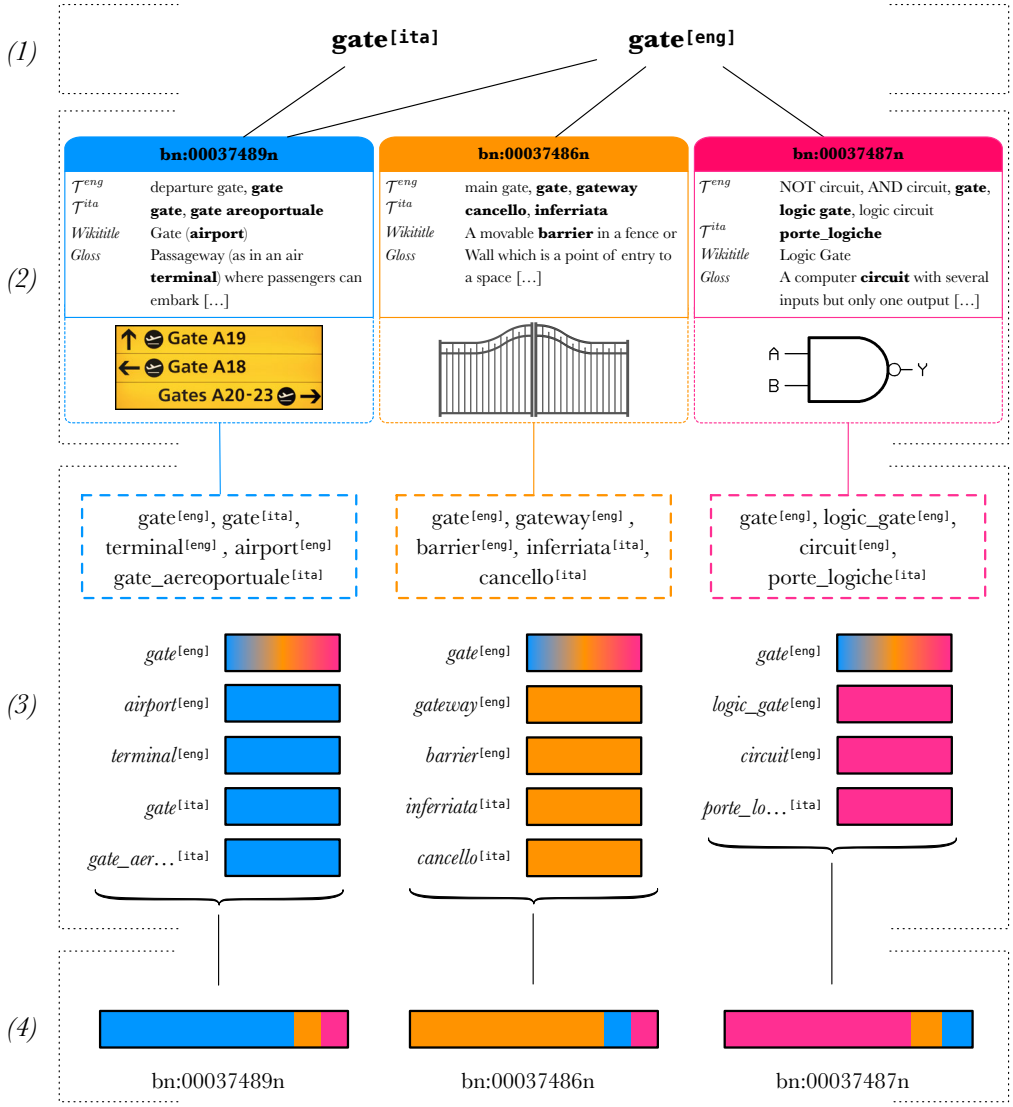

Figure 2
Generation of three LESSLEX vectors, starting from the seed terms $gate^{eng}$ and $gate^{ita}$.

3.3.1 LessLex Building. The LESSLEX resource[8] has been generated from a group of seed terms collected by starting from 56,322 words taken from the Corpus of Contemporary American English (COCA) (Davies 2009),[9] 19,789 terms fetched from the relevant dictionaries of the Internet Dictionary Project,[10] and the 12,544 terms that appear in the data sets that we used during the evaluation. All terms were POS tagged and

8 LESSLEX can be downloaded at the URL `https://ls.di.unito.it/resources/lesslex/`.
9 COCA is a corpus covering different genres, such as spoken, fiction, magazines, newspaper, and academic (`http://corpus.byu.edu/full-text/`).
10 `http://www.june29.com/idp/IDPfiles.html`.

Table 2
Figures on the generation process of LESSLEX, divided by Part of Speech.

LESSLEX Statistics	All	Nouns	Verbs	Adjectives
Seed terms	84,620	45,297	11,943	27,380
Terms in BabelNet	65,629	41,817	8,457	15,355
$\mathcal{T}^+$ avg. cardinality	6.40	6.16	9.67	6.37
Discarded Senses	16,666	14,737	368	1,561
Unique Senses	174,300	148,380	11,038	14,882
Avg. senses per term	4.80	6.12	3.77	1.77
Total extracted terms	227,850	206,603	8,671	12,576
Avg. extracted terms per call	1.40	1.46	1.06	1.05

duplicates removed beforehand. The final figures of the resource and details concerning its generation are reported in Table 2.

We started from a total of 84,620 terms, and for 65,629 of them we were able to retrieve at least one sense in BabelNet. The $\mathcal{T}^+$ cardinality shows that our vectors were built by averaging about six CNN vectors for each concept. Interestingly, verbs seem to have much richer lexical sets. The final number of senses in LESSLEX amounts to 174,300, with a vast majority of nouns. We can also see an interesting overlap between the group of senses associated with each term. If we take nouns, for example, we have around 42*K* terms providing 148*K* unique senses (3.5 per term), while the average polysemy per term counting repetitions amounts to 6.12. So, we can observe that approximately three senses per term are shared with some other term. A large number of concepts are discarded because they only have one term inside $\mathcal{T}^+$: These are named entities or concepts with poor lexicalization sets. The extraction process provided a grand total of about 228*K* terms, and on average each $\mathcal{T}^+$ contains 1.40 additional terms extracted from WTs and glosses.

Out of the 117*K* senses in WordNet (version 3.0), roughly 61*K* of them are covered in LESSLEX. It is, however, important to note that additional LESSLEX vectors can be built upon any set of concepts, provided that they are represented in BabelNet (which contains around 15*M* senses) and that some of their lexicalizations are covered in CNN (1.5*M* terms for the considered languages).

3.3.2 Computing Word Similarity: Maximization and Ranked-Similarity. The word similarity task consists of computing a numerical score that expresses how similar two given terms are. Vectorial resources such as CNN can be easily utilized to solve this task: In fact, because terms are represented as vectors, the distance (usually computed through cosine similarity, or some other variant of angular distance) between the two vectors associated with the input terms can be leveraged to obtain a similarity score. Although terminological resources can be directly used to compute a similarity score between words, conceptually grounded resources (e.g., NASARI, LESSLEX) do not allow us to directly compute world similarity, but rather *conceptual similarity*. In fact, such resources are required to determine which senses must be selected while computing the score for the terms. In most cases this issue is solved by computing the similarity between all the combinations of senses for the two input terms, and then by selecting the maximum

similarity as the result score (Pedersen, Banerjee, and Patwardhan 2005). In formulae, given a term pair $\langle t_1, t_2 \rangle$ and their corresponding list of senses $s(t_1)$ and $s(t_2)$, the similarity can be computed as

$$\text{sim}(t_1, t_2) = \max_{\vec{c_i} \in s(t_1), \vec{c_j} \in s(t_2)} \left[\text{sim}(\vec{c_i}, \vec{c_j}) \right] \tag{1}$$

where $\text{sim}(\vec{c_i}, \vec{c_j})$ is the computation of conceptual similarity using the vector representation for the concepts at hand.

To compute the conceptual similarity between LESSLEX vectors we have devised a different approach, which we call **ranked similarity**. Because we are able to determine not only the distance between each two senses of the input terms, but also the distance between each input term and all of its senses, we use this information to fine tune the computed similarity scores and use ranking as a criterion to grade senses' relevance. In particular, we hypothesize that the relevance of senses for a given term can be helpful for the computation of similarity scores, so we devised a measure that also accounts for the *ranking* of distances between senses and seed term. It implements a heuristics aimed at considering two main elements: the relevance of senses (senses closer to the seed term are preferred), and similarity between sense pairs. Namely, the similarity between two terms t_1, t_2 can be computed as:

$$\begin{aligned} &\text{rnk-sim}(t_1, t_2) \; = \\ &\max_{\vec{c_i} \in s(t_1),\, \vec{c_j} \in s(t_2)} \left[\left((1-\alpha) \cdot (\mathbf{rank}(\vec{c_i}) + \mathbf{rank}(\vec{c_j}))^{-1} \right) + \left(\alpha \cdot \mathbf{cos\text{-}sim}(\vec{c_i}, \vec{c_j}) \right) \right], \end{aligned} \tag{2}$$

where α is used to tune the balance between ranking factor and raw cosine similarity.[11] We illustrate the advantages of the ranked similarity with the following example (Figure 3). Let us consider the two terms *teacher* and *student*, whose gold-standard similarity score is 0.50.[12] One of the senses of teacher is `bn:02193088n` (*The Teacher (1977 film)* - a 1977 Cuban drama film) and one of the senses of student is `bn:02935389n` (*Student (film)* - a 2012 Kazakhstani drama film). These two senses have a cosine similarity in LESSLEX of 0.81; such a high score is reasonable, because they are both drama movies. However, it is clear that an annotator would not refer to these two senses for the input terms, but rather to `bn:00046958n` (*teacher* - a person whose occupation is teaching) and `bn:00029806n` (*student* - a learner who is enrolled in an educational institution). These two senses obtain a similarity score of 0.61, which will not be selected because it is lower than 0.81 (as computed through the formula in Equation (1)). However, if we take into consideration the similarities between the terms *teacher* and *student* and their associated senses, we see that the senses that one would select—while requested to provide a similarity score for the pair—are much closer to the seed terms. The proposed measure involves re-ranking the senses based on their proximity to the term representation, thereby emphasizing more relevant terms. We finally obtain similarity of 0.44 for the movie-related senses, whereas the school-related senses pair obtains a similarity of 0.55, which will be selected and better correlates with human rating.

11 Presently $\alpha = 0.5$.

12 We borrow this word pair from the SemEval 17 Task 2 data set (Camacho-Collados et al. 2017).

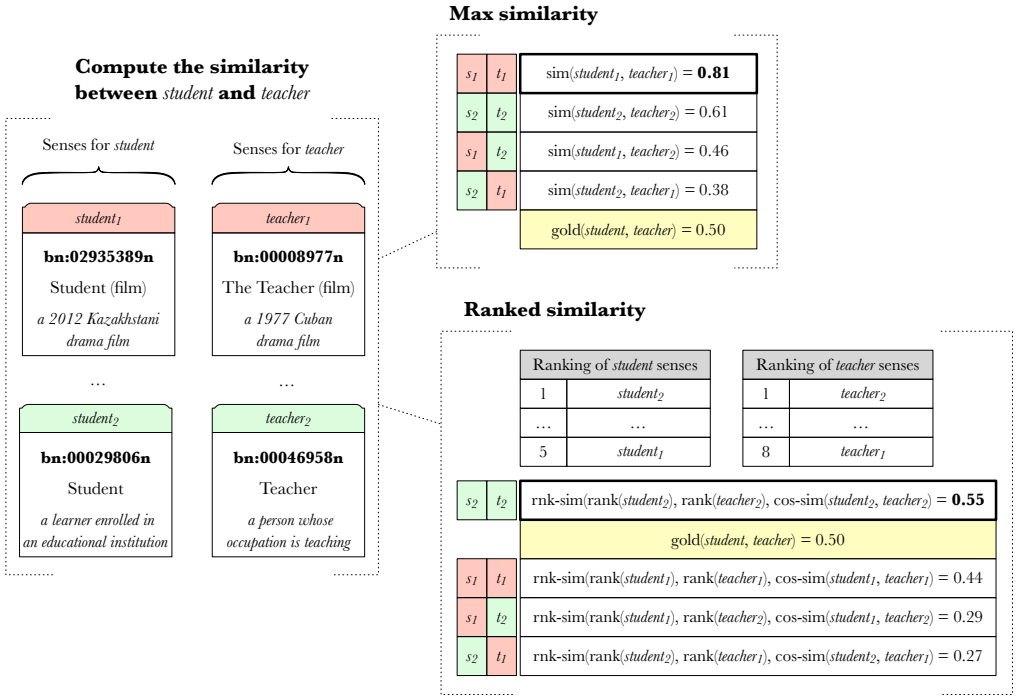

Figure 3
A comparison between the max-similarity (Equation (1)) and the ranked-similarity (Equation (2)) approaches for the computation of the conceptual similarity.

Because the ranked-similarity can be applied only if both terms are available in CNN (so that we can compute the ranks among their senses), we propose a twofold set-up for the usage of LESSLEX. In the first set-up we only make use of the ranked-similarity, so in this setting if at least one given term is not present in CNN we discard the pair as not covered by the resource. In the second set-up (LESSLEX-OOV, designed to deal with *Out Of Vocabulary* terms) we implemented a fallback strategy to ensure higher coverage: In this case, in order to cope with missing vectors in CNN, we adopt the max-similarity as similarity measure in place of the ranked-similarity.

4. Evaluation

In order to assess the flexibility and quality of our embeddings we carried out a set of experiments involving both intrinsic and extrinsic evaluation. Namely, we considered three different tasks:

1. The Semantic Similarity task, where two terms or—less frequently—senses are compared and systems are asked to provide a numerical score expressing how close they are; the systems' output is compared to human ratings (Section 4.1);

2. The more recent Contextual Word Similarity task, asking systems to either assess the semantic similarity of terms taken in context (rather than as pairs of terms taken in isolation), or to decide whether a term has the same meaning in different contexts of usage (Section 4.2); and
3. The Semantic Text Similarity task, where pairs of text excerpts are compared to assess their overall similarity, or to judge whether they convey equal meaning or not (Section 4.3).

4.1 Word Similarity Task

In the first experiment we tested LESSLEX vectors on the word similarity task: Linguistic items are processed in order to compute their similarity, which is then compared against human similarity judgment. Word similarity is mostly thought of as closeness over some metric space, and usually computed through cosine similarity, although different approaches exist, for example, based on cognitively plausible models (Tversky 1977; Jimenez et al. 2013; Lieto, Mensa, and Radicioni 2016a; Mensa, Radicioni, and Lieto 2017). We chose to evaluate our word embeddings on this task because it is a relevant one, for which many applications can be drawn such as Machine Translation (Lavie and Denkowski 2009), Text Summarization (Mohammad and Hirst 2012), and Information Retrieval (Hliaoutakis et al. 2006). Although this is a popular and relevant task, until recently it has been substantially limited to monolingual data, often in English. Conversely, we collected and experimented on all major crosslingual data sets.

4.1.1 Experimental Setting. In this Section we briefly introduce and discuss the selection of data sets adopted for the evaluation.

A pioneering data set is WordSim-353 (Finkelstein et al. 2002); it was built by starting from two older sets of word pairs, the RG-65 and MC-30 data sets (Rubenstein and Goodenough 1965; Miller and Charles 1991). These data sets were originally conceived for the English language and compiled by human experts. They were then translated to multilingual and to crosslingual data sets: The RG-65 has been translated into Farsi and Spanish by Camacho-Collados, Pilehvar, and Navigli (2015a), and the WordSim-353 was translated by Leviant and Reichart (2015b) into Italian, German, and Russian through crowdworkers fluent in such languages. Additionally, WordSim-353 was partitioned by individuating the subset of word pairs appropriate for experimenting on similarity judgments rather than on relatedness judgments (Agirre et al. 2009). The SimLex-999 data set was compiled through crowdsourcing, and includes English word pairs covering different parts of speech, namely, nouns (666 pairs), verbs (222 pairs), and adjectives (111 pairs) (Hill, Reichart, and Korhonen 2015). It has been then translated into German, Italian, and Russian by Leviant and Reichart (2015a). A data set was proposed entirely concerned with English verbs, the SimVerbs-3500 data set (Gerz et al. 2016); similar to SimLex-999, items herein were obtained from the USF free-association database (Nelson, McEvoy, and Schreiber 2004). The SemEval-17 data set was developed by Camacho-Collados et al. (2017); it contains many uncommon entities, like *Si-o-seh pol* or *Mathematical Bridge* encompassing both multilingual and crosslingual data. Finally, another data set was recently released by Goikoetxea, Soroa, and Agirre (2018), in the following referred to as the Goikoetxea data set, built by adding further crosslingual versions for the RG-65, WS-WordSim-353, and SimLex-999 data sets.

In our evaluation both multilingual and crosslingual translations have been used. A *multilingual* data set is one (like RG) where term pairs $\langle x, y \rangle$ from language i have been

translated as $\langle x', y' \rangle$ into a different language, such that both x' and y' belong to the same language. An example is $\langle casa, chiesa \rangle$, $\langle house, church \rangle$, or $\langle maison, église \rangle$. Conversely, in a crosslingual setting (like SemEval 2017, Task 2 - crosslingual subtask), x' is a term from a language different from that of y', like in the pair $\langle casa, church \rangle$.

Many issues can afflict any data set, as is largely acknowledged in the literature (Huang et al. 2012; Camacho-Collados, Pilehvar, and Navigli 2015a; Hill, Reichart, and Korhonen 2015; Camacho-Collados et al. 2017). The oldest data sets are too small (on the order of few tens of word pairs) to attain full statistical significance; until recently, typically similarity and relatedness (association) judgments have been conflated, thereby penalizing models concerned with similarity. Additionally, for such data sets the correlation between systems' results and human ratings is higher than human inter-rater agreement. Because human ratings are largely acknowledged as the upper bound to artificial performance in this kind of task, the point has been raised that such data sets are not fully reliable benchmarks to investigate the correlation between human judgment and systems' output. Furthermore, a tradeoff exists between the size of the data set and the quality of the annotation: Resources acquired through human experts annotation typically are more limited in size, but featured by higher inter-rater agreement (in the order of .80), whereas larger data sets suffer from a lower (often with $< .7$) agreement among annotators, thus implying overall reduced reliability. We thus decided to test on all main data sets adopted in the literature, to provide the most comprehensive evaluation, widening the experimental base as much as possible. The most recent data sets are in principle more controlled and reliable—SimLex-999, SimVerbs, SemEval-2017, Goikoetxea—but still we decided to experiment on all of them, because even RG-65 and WS-Sim 353 have been widely used until recently. All benchmarks used in the experiments are illustrated in Table 3.

The results obtained by using LESSLEX and LESSLEX-OOV are compared with those obtained by utilizing NASARI and CNN, to elaborate on similarities and differences with such resources. Additionally, we report the correlation indices obtained by experimenting with other word and sense embeddings that either are trained to perform on specific data sets (JOINTCHYCB by Goikoetxea, Soroa, and Agirre [2018]), or that directly compare to our resource, as containing both term-level and sense-level vector descriptions (SENSEEMBED and NASARI2VEC). Table 4 summarizes the considered resources and the algorithm used to compute the semantic similarity. In these respects, we adopted the following rationale. When testing with resources that allow for a combined use of word and sense embeddings we use ranked-similarity[13] (as described in Equation (2)); when testing with sense embeddings we adopt the max similarity/closest senses strategy (Resnik 1995; Budanitsky and Hirst 2006; Pilehvar and Navigli 2015) to select senses; when handling word embeddings we make use of the cosine similarity, by borrowing the same approach as illustrated in Camacho-Collados et al. (2017).[14] In order

13 In the experimentation α was set to 0.5.

14 A clarification must be made about SENSEEMBED. Because in this resource both terminological and sense vectors coexist in the same space, the application of the ranked-similarity would be fitting. However, in SENSEEMBED every sense representation is actually indexed on a pair $\langle term, sense \rangle$, so that different vectors may correspond to a given *sense*. In the ranked-similarity, when computing the distance between a term t and its senses, we retrieve the sense identifiers from BabelNet, so to obtain from SENSEEMBED the corresponding vector representations. Unfortunately, however, most senses s_i returned by BabelNet have no corresponding vector in SENSEEMBED associated with the term t (i.e., indexed as $\langle t, s_i \rangle$). This fact directly implies a reduced coverage, undermining the performances of SENSEEMBED. We then realized that the ranked-similarity is an unfair and inconvenient strategy to test on SENSEEMBED (in that it forces using it to some extent improperly), so we resorted to using the max similarity instead.

Table 3
List of the data set employed in the experimentation, showing the POS involved and the languages available in both monolingual and crosslingual versions.

Data set	Part of Speech	Monolingual	Crosslingual
RG-65[1]	nouns	eng, fas, spa	eng, spa, fas, por, fra, deu
WS-Sim-353[2]	nouns	eng, ita, deu, rus	–
SimLex-999[3]	nouns, verbs, adjectives	eng, ita, deu, rus	–
SimVerbs-3500[4]	verbs	eng	–
SemEval 17[5]	nouns	eng, deu, ita, spa, fas	eng, deu, ita, spa, fas
Goikoetxea[6]	nouns, verbs, adjectives	eus	eng, eus, spa, ita

[1] http://lcl.uniroma1.it/similarity-datasets/, https://www.seas.upenn.edu/~hansens/conceptSim/.
[2] http://www.leviants.com/ira.leviant/MultilingualVSMdata.html.
[3] https://fh295.github.io/simlex.html, http://www.leviants.com/ira.leviant/MultilingualVSMdata.html.
[4] http://people.ds.cam.ac.uk/dsg40/simverb.html.
[5] http://alt.qcri.org/semeval2017/task2/index.php?id=data-and-tools.
[6] http://ixa2.si.ehu.es/ukb/bilingual_embeddings.html.

to provide some insights on the quality of the ranked-similarity, we also experiment on an algorithmic baseline referred to as LL-M (LessLex Most Frequent Sense), where we selected the most frequent sense of the input terms based on the connectivity of the considered sense in BabelNet. The underlying rationale is, in this case, to study how this strategy to pick up senses compares with LessLex vectors, which are built from word embeddings that usually tend to encode the most frequent sense of each word. Finally, in the case of the RG-65 data set concerned with sense labeled pairs (Schwartz and Gomez 2011),[15] we only experimented on sense embeddings, and the similarity scores have been computed through the cosine similarity metrics.

4.1.2 Results. All tables report Pearson and Spearman correlations (denoted by r and ρ, respectively); dashes indicate that a given resource does not deal with the considered input, either because of lacking of sense representation, or because of lacking crosslingual vectors. Similarity values for uncovered pairs were set to the middle point of the similarity scale. Additionally, in Appendix A.1 we report the results obtained by considering only the word pairs covered by all the resources: Such figures are of interest, because they allow examining the results obtained from each resource "in purity," by focusing only on their representational precision. All top scores are marked in bold.

Multilingual/Crosslingual RG-65 Data Set. The results obtained over the multilingual and crosslingual RG-65 data set are illustrated in Table 5. RG-65 includes a multilingual data set and a crosslingual one. With regard to the former, both LessLex and LessLex-OOV obtain analogous correlation with respect to CNN when considering term pairs; LessLex and LessLex-OOV substantially outperform NASARI, SenseEmbed, and NASARI2Vec when considering sense pairs (Schwartz and Gomez 2011). Of course

15 This version of the RG-65 data set has been sense-annotated by two humans with WordNet 3.0 senses.

Table 4
List of the resources considered in the experimentation and the algorithm we employed for the resolution of the word similarity task.

	Description	Algorithm
LL-M	LESSLEX	mf-sense similarity
LL-O	LESSLEX (strategy for handling OOV terms)	ranked-similarity
LLX	LESSLEX	ranked-similarity
CNN[1]	ConceptNet Numberbatch word embeddings	cosine similarity
CNN[2]	NASARI sense embeddings	max similarity
JCH[3]	JOINTCHYB bilingual word embeddings	cosine similarity
SSE[4]	SENSEEMBED sense embeddings	max similarity
N2V[5]	NASARI sense embeddings + Word2Vec word embeddings	ranked-similarity

[1] Speer, Chin, and Havasi (2017) (`http://github.com/commonsense/conceptnet-numberbatch` v. 16.09).
[2] Camacho-Collados, Pilehvar, and Navigli (2016) (`http://lcl.uniroma1.it/nasari/` v. 3.0).
[3] Goikoetxea, Soroa, and Agirre (2018) (`http://ixa2.si.ehu.es/ukb/bilingual_embeddings.html`).
[4] Iacobacci, Pilehvar, and Navigli (2015) (`http://lcl.uniroma1.it/sensembed/`).
[5] Word2Vec embeddings trained on UMBC (`http://lcl.uniroma1.it/nasari/`).

Table 5
Results on the multilingual and crosslingual RG-65 data set, consisting of 65 word pairs. With regard to monolingual correlation scores for the English language, we report results for similarity computed by starting from terms (at *words* level), as well as results with sense identifiers (marked as *senses*). The rest of the results were obtained by using word pairs as input. Reported figures express Pearson (*r*) and Spearman (ρ) correlations.

RG-65	LL-M		LLX		LL-O		CNN		NAS		JCH		SSE		N2V	
	r	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ
Word eng	.64	.59	**.91**	.86	**.91**	.86	**.91**	**.90**	.67	.67	.84	.86	.75	.81	.80	.75
Sense eng	–	–	**.94**	**.91**	**.94**	**.91**	–	–	.81	.76	–	–	.72	.76	.78	.73
fas (N)	.75	.72	.75	.75	.73	.70	**.76**	**.76**	.58	.50	–	–	.66	.66 –	–	
spa (N)	.82	.82	**.93**	**.93**	**.93**	**.93**	**.92**	**.93**	.88	.87	.80	.84	.82	.85	–	–
por-fas (N)	.71	.69	.85	.85	.81	.79	**.87**	**.86**	.52	.62	–	–	.70	.66	–	–
fra-por (N)	.82	.83	.92	**.89**	.92	**.89**	**.93**	.88	.69	.67	–	–	.81	.74	–	–
fra-fas (N)	.73	.72	.84	.84	**.86**	.84	**.86**	**.85**	.47	.58	–	–	.72	.71	–	–
fra-spa (N)	.81	.80	**.93**	**.91**	**.93**	**.91**	**.93**	.89	.79	.82	–	–	.88	.86	–	–
fra-deu (N)	.81	.84	**.90**	**.89**	**.90**	**.89**	.88	.87	.77	.77	–	–	.77	.75	–	–
spa-por (N)	.83	.83	**.93**	**.91**	**.93**	**.91**	**.93**	**.91**	.75	.79	–	–	.79	.79	–	–
spa-fas (N)	.71	.70	**.86**	**.87**	.82	.80	**.86**	.86	.50	.64	–	–	.72	.79	–	–
eng-por (N)	.74	.71	**.94**	**.90**	**.94**	**.90**	.92	**.90**	.78	.77	–	–	.80	.76	–	–
eng-fas (N)	.67	.62	**.86**	.85	.84	.81	**.86**	**.87**	.47	.56	–	–	.73	.71	–	–
eng-fra (N)	.71	.70	**.94**	**.92**	**.94**	**.92**	.92	.91	.76	.73	–	–	.81	.75	–	–
eng-spa (N)	.72	.71	**.93**	**.93**	**.93**	**.93**	**.93**	.92	.85	.85	.83	.86	.80	.85	–	–
eng-deu (N)	.74	.72	**.91**	**.89**	**.91**	**.89**	.89	**.89**	.70	.74	–	–	.76	.80	–	–
deu-por (N)	.87	.84	**.91**	**.87**	**.91**	**.87**	**.91**	**.87**	.73	.76	–	–	.76	.72	–	–
deu-fas (N)	.77	.74	.85	**.85**	.87	.84	.85	.84	.58	.65	–	–	.78	.80	–	–
deu-spa (N)	.84	.85	**.91**	**.90**	**.91**	**.90**	.90	.89	.71	.79	–	–	.79	.80	–	–

CNN is not evaluated in this setting, because it only includes representations for terms. With regard to the latter subset, containing crosslingual files, figures show that both CNN and LESSLEX obtained high correlations, higher than the competing resources providing meaning representations for the considered language pairs.

Multilingual WS-Sim-353 Data Set. The results on the multilingual WS-Sim-353 data set are presented in Table 6. Results on these data differ according to the considered language: Interestingly enough, for the English language, the results computed via LESSLEX are substantially on par with those obtained by using CNN vectors. With regard to the remaining translations of the data set, CNN and LESSLEX achieve the highest correlations also on the Italian, German, and Russian languages. Different from other experimental settings (see, e.g., the RG-65 data set), the differences in correlation are more consistent, with LESSLEX obtaining top correlation scores for Italian and Russian, and CNN for German.

Multilingual SimLex-999 Data Set. The results obtained on the SimLex-999 data set are reported in Table 7. We face here twofold results: With regard to the English and the Italian translation, we recorded better results when using the LESSLEX vectors, with consistent advantage over competitors on English verbs. With regard to English adjectives, the highest correlation was recorded when utilizing the LESSLEX Most Frequent Sense vectors (LL-M column). With regard to Italian, as in the WordSim-353 data set,

Table 6
Results on the WS-Sim-353 data set, where we experimented on the 201 word pairs (out of the overall 353 elements) that are acknowledged as appropriated for computing similarity. Reported figures express Pearson (r) and Spearman (ρ) correlations.

WS-Sim-353	LL-M		LLX		LL-O		CNN		NAS		JCH		SSE		N2V	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
eng (N)	.67	.65	**.78**	.78	**.78**	.78	**.78**	**.79**	.60	.61	.72	.72	.69	.73	.71	.70
ita (N)	.67	.68	.70	.73	**.74**	**.78**	.69	.73	.66	.65	.60	.62	.66	.73	–	–
deu (N)	.73	.71	.63	.68	.76	.77	**.82**	**.81**	.64	.63	–	–	.62	.60	–	–
rus (N)	.72	.70	.64	.62	**.73**	**.75**	.65	.63	.63	.61	–	–	.60	.60	–	–

Table 7
Results on the multilingual SimLex-999, including overall 999 word pairs, with 666 nouns, 222 verbs, and 111 adjectives for the English, Italian, German, and Russian languages. Reported figures express Pearson (r) and Spearman (ρ) correlations.

SimLex-999	LL-M		LLX		LL-O		CNN		NAS		JCH		SSE		N2V	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
eng (N)	.51	.50	**.69**	**.67**	**.69**	**.67**	.66	.63	.40	.38	.55	.53	.52	.49	.46	.43
eng (V)	.62	.56	**.67**	**.65**	**.67**	**.65**	.61	.58	–	–	.51	.50	.54	.49	–	–
eng (A)	**.84**	**.83**	.82	.79	.82	.79	.80	.78	–	–	.63	.62	.55	.51	–	–
eng (*)	.57	.55	**.70**	**.69**	**.70**	**.69**	.67	.65	–	–	.55	.54	.53	.49	–	–
ita (N)	.50	.49	**.66**	**.63**	.64	**.63**	.64	.61	.45	.46	.47	.47	.56	.49	–	–
ita (V)	.58	.52	**.69**	**.63**	**.69**	**.63**	.67	.58	–	–	.54	.47	.54	.44	–	–
ita (A)	.65	.58	**.74**	**.69**	**.74**	**.69**	**.74**	.66	–	–	.39	.30	.57	.47	–	–
ita (*)	.51	.47	**.66**	**.62**	.65	**.62**	**.65**	.61	–	–	.46	.44	.54	.47	–	–
deu (N)	.58	.56	.65	.63	.65	.64	**.66**	**.65**	.41	.42	–	–	.47	.43	–	–
deu (V)	.48	.42	.54	.45	.54	.46	**.63**	**.57**	–	–	–	–	.43	.37	–	–
deu (A)	.66	.63	.66	.65	.69	.68	**.77**	**.75**	–	–	–	–	.43	.26	–	–
deu (*)	.55	.52	.62	.59	.63	.61	**.67**	**.65**	–	–	–	–	.45	.38	–	–
rus (N)	.43	.42	.52	.48	.51	**.50**	**.53**	.48	.20	.22	–	–	.26	.21	–	–
rus (V)	.31	.19	.25	.18	.27	.20	**.60**	**.55**	–	–	–	–	.23	.20	–	–
rus (A)	.25	.26	.25	.25	.27	.28	**.69**	**.69**	–	–	–	–	.04	.04	–	–
rus (*)	.36	.32	.43	.37	.42	.39	**.56**	**.51**	–	–	–	–	.23	.13	–	–

the LESSLEX-OOV strategy obtains correlations with human ratings that are higher or on par with respect to those obtained by using LESSLEX vectors. In the second half of the data set CNN performed better on German and Russian.

SimVerbs-3500 Data Set. Results obtained while testing on the SimVerbs-3500 data set are reported in Table 8. In this case it is straightforward to notice that the results obtained by LESSLEX outperform those by all competitors, with a gain of .05 in Pearson *r*, and .06 in Spearman correlation over CNN, on this large set of 3,500 verb pairs. It was not possible to use NASARI vectors, which only exist for noun senses; also notably, the results obtained by using the baseline (LL-M) strategy outperformed those obtained through SENSEEMBED and NASARI2VEC.

Sem Eval 17 Task 2 Data Set. The figures obtained by experimenting on the "SemEval 17 Task 2: Multilingual and Crosslingual Semantic Word Similarity" data set are provided in Table 9. This benchmark is a multilingual data set including 500 word pairs (nouns only) for monolingual versions, and 888 to 978 word pairs for the crosslingual ones.

These results are overall favorable to LESSLEX in the comparison with CNN and with all other competing resources. Interestingly enough, while running the experiments with CNN vectors we observed even higher correlation scores than those obtained in the SemEval 2017 evaluation campaign (Speer, Chin, and Havasi 2017; Camacho-Collados et al. 2017). At that time, such figures scored highest on all

Table 8
Results on the SimVerbs-3500 data set, containing 3,500 verb pairs. Reported figures express Pearson (*r*) and Spearman (ρ) correlations.

SimVerbs	LL-M		LLX		LL-O		CNN		NAS		JCH		SSE		N2V	
	r	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ
eng (V)	.58	.56	**.67**	**.66**	**.67**	**.66**	.62	.60	–	–	.56	.56	.45	.42	.31	.30

Table 9
Results on the SemEval 17 Task 2 data set, containing 500 noun pairs. Reported figures express Pearson (*r*) and Spearman (ρ) correlations.

SemEval 17	LL-M		LLX		LL-O		CNN		NAS		JCH		SSE		N2V	
	r	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ
eng (N)	.71	.72	**.79**	.80	.77	**.81**	**.79**	.79	.64	.65	.50	.45	.69	.73	.64	.64
deu (N)	**.73**	.72	.69	.68	.71	**.75**	.70	.68	.62	.62	–	–	.60	.61	–	–
ita (N)	.74	.75	.66	.65	**.76**	**.79**	.63	.61	.72	.73	.54	.50	.70	.73	–	–
spa (N)	**.77**	.79	.67	.66	.74	**.80**	.63	.62	.72	.73	.50	.48	.68	.71	–	–
fas (N)	.67	.67	.43	.47	**.72**	**.75**	.39	.35	.54	.53	–	–	.60	.63	–	–
deu-spa (N)	**.76**	.77	.69	.68	.74	**.79**	.66	.64	.54	.55	–	–	.65	.68	–	–
deu-ita (N)	**.75**	.76	.68	.67	**.75**	**.79**	.65	.63	.53	.65	–	–	.62	.62	–	–
eng-deu (N)	**.75**	.75	**.75**	.75	**.75**	**.79**	.74	.73	.51	.62	–	–	.63	.63	-	–
eng-spa (N)	.75	.76	.73	.73	**.76**	**.82**	.70	.70	.66	.70	.46	.44	.59	.61	–	–
eng-ita (N)	.74	.76	.72	.72	**.76**	**.82**	.69	.69	.63	.71	.38	.36	.69	.73	–	–
spa-ita (N)	**.76**	.77	.67	.66	**.76**	**.81**	.63	.61	.65	.72	.41	.39	.59	.61	–	–
deu-fas (N)	.72	.73	.55	.52	**.73**	**.76**	.51	.47	.39	.52	–	–	.63	.65	–	–
spa-fas (N)	.72	.73	.55	.52	**.75**	**.79**	.50	.47	.47	.61	–	–	.66	.70	–	–
fas-ita (N)	.72	.73	.53	.50	**.75**	**.78**	.49	.45	.43	.58	–	–	.66	.69	–	–
eng-fas (N)	.71	.72	.58	.55	**.74**	**.79**	.54	.51	.42	.59	–	–	.67	.70	–	–

multilingual tasks (with the exception of the Farsi language) and on all crosslingual settings (with no exception). To date, with regard to the crosslingual setting, LESSLEX correlations indices are constantly higher than those by competitors, including CNN. We observe that the scores obtained by using the baseline with most frequent senses (LL-M) are always ameliorative with respect to all results obtained by experimenting with NASARI, JOINTCHYCB, SENSEEMBED, and NASARI2VEC (with the only exception of the ρ score obtained by SSE on the English monolingual data set).

Multilingual/Crosslingual Goikoetxea Data Set. The results obtained by testing on the Goikoetxea data set are reported in Table 10. The data set includes new variants for three popular data sets: three crosslingual versions for the RG-65 data set (including the Basque language, marked as "eus" in the table); the six crosslingual combinations of the Basque, Italian, and Spanish translations of the WS-Sim-353 data set; and three crosslingual translations of the SimLex-999 data set, including its English, Italian, and Spanish translations.

Results are thus threefold. With regard to the first block on the RG-65 data set, LESSLEX results outperform all competitors (to a smaller extent on versions involving the Basque language), including JOINTCHYCB, the best model by Goikoetxea, Soroa, and Agirre (2018). In the comparison with CNN, LESSLEX vectors achieve better results, with higher correlation for cases involving Basque, on par on the English–Spanish data set. With regard to the second block (composed of crosslingual translations of the WS-Sim-353 data set), we record that the LESSLEX-OOV strategy obtained the top Spearman correlation scores, coupled with poor Pearson correlation scores; whereas CNN and JCH obtain the best results with regard to the latter coefficients. In the last

Table 10
Results on the Goikoetxea data set. The data set includes variants of the RG-65 (first block), WS-Sim-353 (second block) and SimLex-999 (third block) data sets. The "eus" abbreviation indicates the Basque language. Reported figures express Pearson (r) and Spearman (ρ) correlations.

Goikoetxea	LL-M		LLX		LL-O		CNN		NAS		JCH		SSE		N2V	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
spa-eus (N)	.74	.72	.42	.67	**.76**	**.77**	.66	.61	.71	.74	.73	.72	.61	.71	–	–
eng-eus (N)	.74	.74	.41	.77	**.89**	**.91**	.77	.73	**.89**	.88	.88	.87	.81	.83	–	–
eng-spa (N)	.72	.71	**.93**	**.93**	**.93**	**.93**	**.93**	**.93**	.77	.82	.83	.86	.64	.85	–	–
eus-ita (N)	.27	.68	.42	**.74**	.24	.71	.51	.53	.49	.56	**.52**	.58	.20	.58	–	–
spa-ita (N)	.29	.66	.29	**.76**	.29	.74	**.63**	.70	.53	.57	.54	.60	.21	.59	–	–
spa-eus (N)	.31	.74	.40	**.78**	.29	**.78**	.55	.56	.59	.66	**.69**	.73	.23	.64	–	–
eng-ita (N)	.30	.64	.27	**.77**	.32	.76	**.67**	.74	.47	.52	.59	.64	.21	.59	–	–
eng-eus (N)	.30	.70	.39	**.79**	.29	.78	.56	.57	.52	.60	**.71**	.75	.23	.64	–	–
eng-spa (N)	.34	.66	.27	**.79**	.40	.77	**.70**	.76	.52	.56	.68	.73	.29	.64	–	–
eng-spa (N)	.49	.48	**.66**	**.64**	.65	**.64**	.64	.62	.36	.46	.54	.51	.53	.50	–	–
eng-spa (V)	.54	.50	.61	.59	**.62**	**.60**	.58	.56	–	–	.43	.43	.52	.49	–	–
eng-spa (A)	.72	.73	.73	.74	.72	**.75**	**.74**	.74	–	–	.56	.55	.53	.47	–	–
eng-spa (*)	.53	.51	**.66**	.64	.65	**.65**	.64	.63	–	–	.50	.52	.53	.49	–	–
eng-ita (N)	.52	.52	**.70**	**.68**	**.70**	**.68**	.68	.66	.36	.45	.51	.50	.54	.51	–	–
eng-ita (V)	.49	.40	.57	.51	.57	.51	**.67**	**.62**	–	–	.47	.51	.44	.33	–	–
eng-ita (A)	.75	.74	**.79**	**.78**	**.79**	**.78**	.77	.72	–	–	.42	.43	.57	.45	–	–
eng-ita (*)	.50	.46	.65	.62	.65	.63	**.68**	**.66**	–	–	.48	.50	.51	.43	–	–
spa-ita (N)	.53	.53	**.67**	.65	**.67**	**.66**	.66	.64	.34	.45	.45	.45	.54	.52	–	–
spa-ita (V)	.44	.39	.51	.46	.51	.46	**.63**	**.60**	–	–	.42	.44	.43	.34	–	–
spa-ita (A)	.68	.66	.73	.71	.72	**.73**	**.73**	.69	–	–	.41	.45	.57	.48	–	–
spa-ita (*)	.49	.46	.61	.58	.61	.59	**.66**	**.64**	–	–	.44	.45	.50	.45	–	–

block of results in Table 10 (containing translations for the SimLex-999 data set), we first observe that comparing the obtained figures is not simple: We report the figures obtained by Goikoetxea, Soroa, and Agirre (2018) with no distinction in POS. However, if we focus on results on nouns (two thirds of the SimLex-999 data set), LessLex vectors obtain the best results, although it is not easy to determine whether LessLex or CNN vectors provided the overall best results on the other parts of speech.

4.1.3 Discussion. We overall experimented on nine different languages (deu, eng, eus, fas, fra, ita, por, rus, spa) and various crosslingual combinations. Collectively, such tests constitute a widely varied experimental setting, to the best of our knowledge the largest on the semantic similarity task. The obtained results allow us to state that LessLex is at least on par with competing state-of-the-art resources, although we also noticed that some room still exists for further improvements, such as the coverage on individual languages (e.g., Russian and German).

Let us start by considering the results on the multilingual WS-Sim-353 and on the SimLex data sets (Tables 6 and 7, respectively). The results obtained through LessLex always improve on those obtained by using the sense embeddings by SenseEmbed and NASARI2Vec, which provide term and sense descriptions embedded in the same semantic space, and are thus closer to our resource. Also, the comparison with NASARI is favorable to LessLex. In the comparison with CNN, we note that whereas in the English language LessLex and LessLex-OOV scores either outperform or closely approach those obtained through CNN, in other languages our vectors suffer from the reduced and less rich sense inventory of BabelNet, which in turn determines a lower quality for our vectors. This can be easily identified if one considers that a less rich synset contains fewer terms to be plugged into our vectors, thereby determining an overall poorer semantic coverage. The poor results obtained by utilizing LessLex on the German and Russian subsets of the WS-Sim-353 and SimLex-999 data sets probably stem from this sort of limitation.

A consistent difference between LessLex ranked-similarity and the LessLex-OOV strategy can be observed when a sense is available in BabelNet, but not the corresponding vector in CNN: The LessLex-OOV strategy basically consists of resorting to the maximization approach when—due to the lack of a terminological description associated with the sense at hand—it is not possible to compute the ranked-similarity. This strategy was executed in around 9% of cases ($\sigma = 12\%$) over all data sets, ranging from 0% on verbs in the SimVerbs-3500 data set, up to around 50% for the Farsi nouns in the SemEval-2017 monolingual data set. Although not used often, this strategy contributed in many cases to obtain top scoring results, improving on those computed with plain ranked-similarity with LessLex, and also in some cases on CNN and NASARI, as illustrated in both the monolingual and crosslingual portions of the SemEval-2017 data set (Table 9).

Cases where results obtained through LessLex improve over those obtained with CNN are important to assess LessLex, in that they confirm that the control strategy for building our vectors is effective, and that our vectors contain precise and high-quality semantic descriptions. In this sense, obtaining higher or comparable results by using sense embeddings with respect to using word embeddings (with sense embeddings featuring an increased problem space with respect to the latter ones) is per se an achievement. Additionally, our vectors are grounded on BabelNet synset identifiers, which allows us to address each sense as part of a large semantic network, providing further information on senses with respect to the meaning descriptions conveyed through the 300-dimensional vectors. While the LessLex-OOV is a run-time strategy concerned with

the usage of LESSLEX to compare sense pairs, the quality of our vectors is determined by the enrichment step. More specifically, the coverage of our vectors depends on the strategy devised to build $\mathcal{T}^+$ because the coverage is determined both by the number of term-level vectors, and by the number of sense vectors associated with each term, so that in a sense the coverage of LESSLEX is determined by the size of $\mathcal{T}^+$. Additionally, we register that the elements added to the extended set $\mathcal{T}^+$ are often of high quality, as proven, for example, by the sense-oriented task of the RG-65 data set, where senses were assessed (Table 5, line 2): In this setting, the correlation indices for LESSLEX and LESSLEX-OOV vectors score highest over all semantic resources, including NASARI, SENSEEMBED, and NASARI2VEC.

Additionaly results achieved while testing on the Goikoetxea data set seem to confirm that our LL-O strategy allows us to deal with languages with reduced (with respect to English) coverage and/or sense inventory in either BabelNet or ConceptNet: In 12 out of the overall 18 tests on this data set, the LESSLEX-OOV strategy earned at least one top scoring correlation index (either *r* or ρ, as shown in Table 10). The comparison with the recent JOINTCHYCB embeddings shows that the adoption of a shared conceptual—multilingual—level can be beneficial and advantageous with respect to building specialized pairs of embeddings.

Less relevant under a crosslingual perspective, but perhaps relevant in order to fully assess the strengths of our resource, LESSLEX vectors achieved by far the highest correlation scores on English verbs (refer to Table 7, line 2, and Table 8). The comparison with previous literature seems to corroborate this fact (Gerz et al. 2016): In fact, to the best of our knowledge previous state-of-the-art systems achieved around .624 Spearman correlation (Mrkšić et al. 2016; Faruqui and Dyer 2015).

In order to further deepen the analysis of the results, it is instructive to compare the results reported in Tables 5–10 with those obtained on the fraction of data set covered by all considered resources, and provided in Appendix A (Tables 17–22). That is, for each data set we re-run the experiments for all considered resources by restricting to compare only term pairs actually covered by all resources. We call this evaluation metric *CbA condition* hereafter (from "Covered by All"); as opposed to the case in which a mid-scale similarity value was assigned to uncovered terms, referred to as the *MSV condition* in the following (from "Mid Scale Value"). As mentioned, the CbA condition allows evaluating the representational precision of the resources at stake independent of their coverage, whereas a mixture of both aspects is grasped in the the MSV condition. In the leftmost column of the tables in Appendix A we report the coverage for each test. As we can see, coverage is diverse across data sets, ranging from .61 (averaged on all variants, with a minimum on the Farsi language, in the order of .34 and all translations involving the Farsi) in the SemEval-2017 data set (Table 21) to 1.0 in the SimVerbs-3500 data set (Table 19). Other notable cases in which relevant variations in coverage were observed are Russian verbs and adjectives in the SimLex-999 data set, with .20 and .06 coverage, respectively (Table 20). In general, as expected, the recorded correlations are improved with respect to results registered for the corresponding (same data set and resource) test in the MSV set-up, although spot pejorative cases were observed, as well (see, e.g., CNN results for Italian adjectives, in the SimLex-999 data set, reported in Table 20). For example, if we consider the poorly covered SemEval-2017 data set, we observe the following rough improvements (average over all translations, and both *r* and ρ metrics) in the correlation indices: .20 for LESSLEX, .22 for CNN, .09 for NASARI, .30 for JOINTCHYCB (that does not cover all translations, anyway), .07 for SENSEEMBED, and .09 for NASARI2VEC (only dealing with nouns).

Table 11
The top half of the table shows a synthesis of the results obtained in the Mid-Scale similarity Value (MSV) experimental condition, whose details have been illustrated in Tables 5–10; at the bottom we provide a synthesis of the results obtained in the Covered by All (CbA) experimental condition, illustrated in detail in Tables 17–22.

Mid-Scale similarity Value (MSV) Experimental Condition								
	LL-M	LLX	LL-O	CNN	NAS	JCH	SSE	N2V
Spearman ρ	7	32	41	33	1	3	0	0
Pearson *r*	1	32	50	24	0	0	0	0
Total	8	64	91	57	1	3	0	0

Covered by All (CbA) Experimental Condition								
	LL-M	LLX	LL-O	CNN	NAS	JCH	SSE	N2V
Spearman ρ	1	61	–	30	0	0	0	0
Pearson *r*	2	63	–	22	0	0	0	0
Total	3	124	–	52	0	0	0	0

In order to synthetically examine how the CbA experimental condition affected results with respect to the MSV condition, we adopt a rough index, simply counting the number of test results (we consider as a separate test result each Pearson and each Spearman score in Tables 17–22) where each resource obtained highest scores.[16] We thus count overall 152 tests (15 in the SemEval-2017 data set, 4 in the WS-Sim-353, 1 in the SimVerbs-3500, 16 in the SimLex-999, 19 in the RG-65, and 21 in the Goikoetxea; for each one we consider as separated *r* and ρ scores). Provided that in several cases we recorded more than one single resource attaining top scores, the impact of the reduced coverage (CbA condition) vs. MSV condition is presented in Table 11. In the MSV condition we have LESSLEX-OOV achieving 91 top scoring results, followed by LESSLEX with 64 and CNN with 57. In the CbA experimental condition, the LESSLEX-OOV strategy was never executed (because only the actual coverage of all resources was considered, and no strategy for handling out-of-vocabulary terms was thus necessary), and LESSLEX obtained 124 top scoring results, against 52 for CNN. In the latter condition there were fewer cases with a tie. All in all, we interpret the different correlation scores obtained in the two experimental conditions as an evidence that LESSLEX embeddings are featured by good coverage (as suggested by the results obtained in the MSV condition) and lexical precision (as suggested by the results obtained in the CbA condition), improving on those provided by all other resources at stake.

Our approach showed to scale well to all considered languages, under the mild assumption that these are covered by BabelNet, and available in the adopted vectorial resource; when such conditions are met, LESSLEX vectors can be in principle built on a streamlined, on-demand, basis, for any language and any POS.

16 Of course we are aware that this is only a rough index, which, for example does not account for the data sets size (varying from 65 to 3,500 word pairs) or the involved POS, and mixing Pearson and Spearman correlation scores.

Table 12
Some descriptive statistics of the WiC data set. In particular, the distribution of nouns and verbs, number of instances and unique words across training, development and test set of the WiC data set are reported.

Split	Instances	Nouns	Verbs	Unique Words
Training	5,428	49%	51%	1,256
Dev	638	62%	38%	599
Test	1,400	59%	41%	1,184

4.2 Contextual Word Similarity Task

As the second test bed we experimented on the contextual word similarity task, which is a variant of the word similarity. In this scenario the target words are taken *in context*, meaning that the input word is given as input together with the piece of text in which they occur. In this setting, systems are required to account for meaning variations in the considered context, so that typical static word embeddings such as Word2Vec, ConceptNet Numberbatch, and so forth, are not able to grasp their mutable, dynamic semantics. We tested on both Stanford's Contextual Word Similarities (SCWS) data set (Huang et al. 2012), and on the more recent Word-in-Context (WiC) data set (Pilehvar and Camacho-Collados 2019). The SCWS data set defines the problem as a similarity task, where each input record contains two sentences in which two distinct target words t_1 and t_2 are used. The task requires providing the pair $\langle t_1, t_2 \rangle$ with a similarity score by taking into account the context where the given terms occur. The data set consists of 2,003 instances, divided into 1,328 instances whose targets are a noun pair, 399 a verb pair, 97 an adjectival pair, 140 contain a verb-noun pair, 30 contain a noun-adjective pair, and 9 a verb-adjective pair. On the other hand, in the WiC data set the contextual word similarity problem is cast to a binary classification task: Each instance is composed of two sentences in which a specific target word t is used. The utilized algorithm has to make a decision on whether t assumes the same meaning or not in the two given sentences. The distribution of nouns and verbs across training, development, and test-set is reported in Table 12, together with figures on number of instances and unique words.

In the following we report the results obtained on the two data sets by experimenting with LESSLEX and the ranked-similarity metrics. Our results are compared with those reported in literature, and with those obtained by experimenting with NASARI2VEC, which is the only competing resource suitable to implement the ranked similarity along with its contextual variant.

4.2.1 Testing on the SCWS Data Set. To test on the SCWS data set we used both the ranked-similarity (rnk-sim) and the *contextual* ranked-similarity (c-rnk-sim), a variant devised to account for contextual information. With regard to the latter one, given two sentences $\langle S_1, S_2 \rangle$, we first computed the context vectors $\langle \overrightarrow{ctx_1}, \overrightarrow{ctx_2} \rangle$ with a bag-of-words approach, that is, by averaging all the terminological vectors of the lexical items contained therein:

$$\overrightarrow{ctx_i} = \frac{\sum_{t \in S_i} \vec{t}}{N} \tag{3}$$

where N is the number of words in the sentence S_i.

Table 13
Results obtained by experimenting on the SCWS data set. Figures report the *Spearman* correlations with the gold standard divided by part of speech. In the top of the table we report our own experimental results, while, in the bottom, results from literature are provided.

System	ALL	N-N	N-V	N-A	V-V	V-A	A-A
LESSLEX (rnk-sim)	0.695	0.692	0.696	0.820	0.641	0.736	0.638
LESSLEX (c-rnk-sim)	0.667	0.665	0.684	0.744	0.643	0.725	0.524
NASARI2VEC (rnk-sim)	–	0.384	–	–	–	–	–
NASARI2VEC (c-rnk-sim)	–	0.471	–	–	–	–	–
SENSEEMBED[1]	0.624	–	–	–	–	–	–
Huang et al. 50d[2]	0.657	–	–	–	–	–	–
Arora at al.[3]	0.652	–	–	–	–	–	–
MSSG.300D.6K[4]	0.679	–	–	–	–	–	–
MSSG.300D.30K[4]	0.678	–	–	–	–	–	–

[1] Iacobacci, Pilehvar, and Navigli (2015).
[2] Huang et al. (2012).
[3] Arora et al. (2018).
[4] Neelakantan et al. (2014), figures reported from Mu, Bhat, and Viswanath (2017).

The two context vectors are then used to perform the sense rankings for the target words, in the same fashion as in the original ranked-similarity:

$$\text{c-rnk-sim}(t_1, t_2, \overrightarrow{ctx_1}, \overrightarrow{ctx_2}) = \max_{\substack{\vec{c_i} \in s(t_1) \\ \vec{c_j} \in s(t_2)}} \left[\left((1-\alpha) \cdot \left(\underbrace{\text{rank}(\vec{c_i})}_{\text{w.r.t. } \overrightarrow{ctx_1}} + \underbrace{\text{rank}(\vec{c_j})}_{\text{w.r.t. } \overrightarrow{ctx_2}} \right)^{-1} \right) + \left(\alpha \cdot \text{cos-sim}(\vec{c_i}, \vec{c_j}) \right) \right] \quad (4)$$

Results. The results obtained by experimenting on the SCWS data set are reported in Table 13.[17] In spite of the simplicity of the system using LESSLEX embeddings, our results overcome those reported in literature, where by far more complex architectures were used.

However, such scores are higher than the agreement among human raters, which can be thought of as an upper bound to systems' performance. The Spearman correlation among human ratings (computed on leave-one-out basis, that is, by averaging the correlations between each rater and the average of all others) is reportedly 0.52 for the SCWS data set (Chi and Chen 2018; Chi, Shih, and Chen 2018), which can be considered as a poor inter-rater agreement. Also to some extent surprising is the fact that the simple ranked-similarity (rnk-sim), which was intended as a plain baseline, surpassed the contextual ranked-similarity (c-rnk-sim), more suited for this task.

To further elaborate on our results we then re-ran the experiment by investigating how the obtained correlations are affected by different degrees of consistency in the annotation. We partitioned the data set items based on the standard deviation recorded in human ratings, obtaining 9 bins, and re-ran our system on these, utilizing both metrics, with the same parameter settings as in the previous run. In this case the Pearson

17 Parameters setting: in rnk-sim and in the c-rnk-sim α was set to 0.5 for both LESSLEX and NASARI2VEC.

Table 14
Correlation scores obtained with LESSLEX on different subsets of data obtained by varying standard deviation in human ratings. The reported figures show higher correlation when testing on the most reliable (with smaller standard deviation) portions of the data set. To interpret the standard deviation values, we recall that the original ratings collected in the SCWS data set were expressed in the range [0.0, 10.0].

σ	c-rank-sim (*r*)	rank-sim (*r*)	nof-items
≤ 0.5	0.83	0.82	39
≤ 1.0	0.85	0.86	82
≤ 1.5	0.85	0.85	165
≤ 2.0	0.82	0.84	285
≤ 2.5	0.68	0.83	518
≤ 3.0	0.68	0.79	903
≤ 3.5	0.67	0.75	1,429
≤ 4.0	0.64	0.71	1,822
< 5.0	0.63	0.69	2,003

correlation indices were recorded, in order to investigate the linear relationship between our output and human ratings. As expected, we obtained higher correlations on the most reliable portions of the data set, those with smallest standard deviation (Table 14).

However, we still found surprising the obtained results, since the rnk-sim metric seems to be more robust than its contextual counterpart. This is in contrast with literature, where the top scoring metrics, originally defined by Reisinger and Mooney (2010), also leverage contextual information (Huang et al. 2012; Chen, Liu, and Sun 2014; Chen et al. 2015). In particular, the *AvgSim* metrics (which is computed as a function of the average similarity of all prototype pairs, without taking into account the context) is reportedly outperformed by the *AvgSimC* metrics, in which terms are weighted by the likelihood of the word contexts appearing in the respective clusters). The *AvgSim* and the *AvgSimC* directly compare to our rnk-sim and c-rnk-sim metrics, respectively. In our results, for the lowest levels of standard deviation (that is, for $\sigma \leq 2$), the two metrics perform in a similar way; for growing values of σ we observe a substantial drop of the c-rank-sim, while the correlation of the rnk-sim decreases more smoothly. In these cases (for $\sigma \geq 2.5$) contextual information seems to be less relevant than pair-wise similarity of term pairs taken in isolation.

4.2.2 Testing on the WiC Data Set. Different from the SCWS data set, in experimenting on WiC we are required to decide whether a given term conveys the same or different meaning in their context, as in a binary classification task. Context-insensitive word embedding models are expected here to approach a random baseline, while the upper bound, provided by human-level performance, is 80% accuracy.

We run two experiments, one where the contextual ranked-similarity was used, the other with the Rank-Biased Overlap (RBO) (Webber, Moffat, and Zobel 2010). In the former case, we used the *contextual* ranked-similarity (Equation (4)) as the metrics to compute the similarity score, and we added a similarity threshold to provide a binary answer. In the latter case, we designed another simple schema to assess the semantic similarity between term senses and context. At first we built a context vector (Equation (3)) to acquire a compact vectorial description of both texts at hand, obtaining two context vectors $\overrightarrow{ctx_1}$ and $\overrightarrow{ctx_2}$. We then ranked all senses of the term of interest (based

on the cosine similarity metrics) with respect to both context vectors, obtaining s_1^t and s_2^t, as the similarity ranking of t senses from $\overrightarrow{ctx_1}$ and $\overrightarrow{ctx_2}$, respectively. The RBO metrics were then used to compare the similarity between such rankings. Given two rankings s_1^t and s_2^t, RBO is defined as follows:

$$\text{RBO}(s_1^t, s_2^t) = (1-p)\sum_{d=1}^{|O|} p^{d-1}\frac{|O_d|}{d} \tag{5}$$

where O is the set of overlapping elements, $|O_d|$ counts the number of overlaps out of the first d elements, and p is a parameter governing how steep the decline in weights is—setting p to 0 would imply considering only the top element of the rank. In this setting, a low RBO score can be interpreted as indicating that senses that are closest to the contexts are different (thus suggesting that the sense intended by the polysemous term is different across texts), whereas the opposite case indicates that the senses more fitting to both contexts are the same or similar, thereby authorizing judging them as similar. For the task at hand, we simply assigned same sense when the RBO score exceeded a threshold set to 0.8.[18]

Results. The results obtained experimenting on the WiC data set are reported in Table 15. Previous results show that this data set is very challenging for embeddings that do not directly grasp contextual information. The results of systems participating in this task can then be arranged into three main classes: those adopting embeddings featured by contextualized word embeddings, those experimenting with embeddings endowed with sense representations, and those implementing sentence-level baselines (Pilehvar and Camacho-Collados 2019). Given that the data set is balanced (that is, it comprises an equal number of cases where the meaning of the polysemous term is preserved/different across sentences), and the fact that the task is a binary classification one, the random baseline is 50% accuracy. Systems utilizing sense representations (directly comparing to ours) obtained up to 58.7% accuracy score (Pilehvar and Collier 2016. On the other side, those using contextualized word embeddings achieved accuracy ranging from 57.7% accuracy (ELMo 1024-*d*, from the first LSTM hidden state) to 68.4% accuracy (BERT 1024-*d*, 24 layers, 340M parameters) (Pilehvar and Camacho-Collados 2019).

Our resource directly compares with multi-prototype, sense-oriented, embeddings, namely, JBT (Pelevina et al. 2016), DeConf (Pilehvar and Collier 2016), and SW2V (Mancini et al. 2017). In spite of the simplicity of both adopted approaches (c-rnk-sim and RBO), by using LESSLEX vectors we obtained higher accuracy values than those reported for such comparable resources (listed as "Sense representations" in Table 15).

We also experimented with N2V (with both c-rank-sim and RBO metrics), whose results are reported for nouns on the training and development subsets.[19] For such partial results we found slightly higher accuracy than obtained with LESSLEX with the RBO metrics. Unfortunately, however, N2V results can hardly be compared to ours, because the experiments on the test set were executed through the CodaLab

18 The RBO parameter p has been optimized and set to .9, which is a setting also in accord with the literature (Webber, Moffat, and Zobel 2010).

19 Parameters setting for NASARI2VEC: in the c-rnk-sim, α was set to 0.7, and the threshold to 0.8; in the RBO run, p was set to 0.9 and the threshold to 0.9.

Table 15
Results obtained by experimenting on the WiC data set. Figures report the accuracy obtained for the three portions of the data set and divided by POS.

System	Test	Training			Development		
		All	Nouns	Verbs	All	Nouns	Verbs
Contextualized word embeddings							
BERT-large[1]	68.4	–	–	–	–	–	–
WSD[2]	67.7	–	–	–	–	–	–
Ensemble[3]	66.7	–	–	–	–	–	
BERT-large[4]	65.5	–	–	–	–	–	–
ELMo-weighted[5]	61.2	–	–	–	–	–	–
Context2vec[4]	59.3	–	–	–	–	–	–
Elmo[4]	57.7	–	–	–	–	–	–
Sense representations							
DeConf[4]	58.7	–	–	–	–	–	–
SW2V[4]	58.1	–	–	–	–	–	–
JBT[4]	53.6	–	–	–	–	–	
LessLex (c-rnk-sim)	58.9	59.4	58.8	60.1	60.5	58.0	64.6
LessLex (RBO)	59.2	61.1	59.4	62.9	63.0	62.0	64.6
N2V (c-rnk-sim)	–	–	54.1	–	–	53.2	–
N2V (RBO)	–	–	60.7	–	–	63.4	–

[1] Wang et al. (2019).
[2] Loureiro and Jorge (2019).
[3] Soler, Apidianaki, and Allauzen (2019).
[4] Mancini et al. (2017).
[5] Ansell, Bravo-Marquez, and Pfahringer (2019).

Competitions framework.[20] In fact, the design of the competition does not permit us to separate the results for nouns and verbs, as the gold standard for the test set is not publicly available,[21] so we were not able to directly experiment on the test set to deepen comparisons.

4.3 Semantic Text Similarity Task

As our third and final evaluation we consider the *Semantic Text Similarity* (STS) task, an extrinsic task that consists in computing a similarity score between two given portions of text. STS plays an important role in a plethora of applications such as information retrieval, text classification, question answering, topic detection, and as such it is helpful to evaluate to what extent LessLex vectors are suited to a downstream application.

Experimental Set-Up. We provide our results on two data sets popular for this task: the STS benchmark, and the SemEval-2017 Task 1 data set, both by Cer et al. (2017).

20 `https://competitions.codalab.org/competitions/20010`.
21 As of mid August 2019.

The former data set has been built by starting from the corpus of English SemEval STS shared task data (2012–2017). Sentence pairs in the SemEval-2017 data set feature a varied crosslingual and multilingual setting, deriving from the Stanford Natural Language for Inference (Bowman et al. 2015) except for one track (one of two Spanish–English crosslingual tasks, referred to as Track 4b. spa-spa), whose linguistic material has been taken from the WMT 2014 quality estimation task by Bojar et al. (2014). The translations in this data set are the following: Arabic (ara-ara), Arabic-English (ara-eng), Spanish (spa-spa), Spanish-English (spa-eng), Spanish-English (spa-eng), English (eng-eng), Turkish-English (tur-eng).

To assess our embeddings in this task, we used the implementation of the HCTI system, participating in the SemEval-2017 Task 1 (Shao 2017), kindly made available by the author.[22] HCTI obtained the overall third place in that SemEval competition. The HCTI system—implemented by using Keras (Chollet 2015) and Tensorflow (Abadi et al. 2016)—generates sentence embeddings with twin convolutional neural networks; these are then compared through the cosine similarity metrics, and element-wise difference with the resulting values is fed to additional layers to predict similarity labels. Namely, a Fully Connected Neural Network is used to transfer the semantic difference vector to a probability distribution over similarity scores. Two layers are used herein, the first one using 300 units with *tanh* activation function; the second layer is charged to compute the (similarity label) probability distribution with 6 units combined with *softmax* activation function. Whereas the original HCTI system uses GloVe vectors (Pennington, Socher, and Manning 2014), we used LESSLEX vectors in our experimentation.

In order to actually compare only the utilized vectors by leaving unaltered the rest of the HCTI system, we adopted the same parameter setting as is available in the software bundle implementing the approach proposed in Shao (2017). We were basically able to reproduce the results of the paper, except for the hand-crafted features; however, based on experimental evidence, these did not seem to produce significant improvements in the system's accuracy.

We devised two simple strategies to choose the word-senses to be actually fed to the HCTI system. In the first case we built the context vector (as illustrated in Equation (3)), and selected for each input term the sense closest to such vector. The same procedure has been run on both texts being compared for similarity. In the following we refer to this strategy as *c-rank*. In the second case we selected for each input term the sense closest to the terminological vector, in the same spirit as in the first component of the ranked similarity (rnk-sim, Equation (2)). In the following this strategy is referred to as *t-rank*.

As mentioned, in the original experimentation two runs of the HCTI system were performed: one exploiting MT to translate all sentences into English, and another one with no MT, but performing a specific training on each track, depending on the involved languages (Shao 2017, page 132). Because we are primarily interested in comparing LESSLEX and GloVe vectors, rather than the quality of services for MT, we experimented in the condition with no MT. However, in this setting the GloVe vectors could not be directly used to deal with the crosslingual tracks of the SemEval-2017 data set. Specific retraining (although with no handcrafted features) was performed by the HCTI system using the GloVe vectors on the multilingual tracks. In experimenting with LESSLEX vectors, the HCTI system was trained only on the English STS benchmark data set also

22 `http://tiny.cc/dstsaz`.

Table 16
Results on the STS task. Top: results on the STS benchmark. Bottom: results on the SemEval-2017 data set. Reported results are Pearson correlation indices, measuring the agreement with human annotated data. In particular, we compare the Pearson scores obtained by the HCTI system using LESSLEX and GloVe vectors. With regard to the runs with GloVe vectors, we report results with no hand-crafted features (no HF), and without machine translation (no MT).

STS Benchmark (English)			
Track	HCTI + LESSLEX		HCTI + GloVe
	(t-rank)	(c-rank)	(no HF)
dev	.819	.823	.824
test	.772	.786	.783
SemEval 2017			
Track	HCTI + LESSLEX		HCTI + GloVe
	(t-rank)	(c-rank)	(no MT)
1. ara-ara	.534	**.618**	.437
2. ara-eng	.310	.476	–
3. spa-spa	**.800**	.730	.671
4a. spa-eng	.576	.558	–
4b. spa-eng	.143	.009	–
5. eng-eng	.811	.708	**.816**
6. tur-eng	.400	.433	–

to deal with the SemEval-2017 data set: that is, no MT step nor any specific re-training was performed in experiments with LESSLEX vectors to deal with crosslingual tracks.

Results. Results are reported in Table 16, where the correlation scores obtained by experimenting with LESSLEX and GloVe vectors are compared.

Let us start by considering the results obtained by experimenting on the STS benchmark. Here, when using LESSLEX embeddings we obtained figures similar to those obtained by the HCTI system using GLoVe vectors; namely, we observe that the choice of senses based on the overall context (c-rank) provides little improvement with respect to both GloVe vectors and to the t-rank strategy.

With regard to the seven tracks in the SemEval-2017 data set, we can distinguish between results on multilingual and crosslingual subsets of data. With regard to the former ones (that is, the ara-ara, spa-spa, and eng-eng tracks), HCTI with LESSLEX obtained higher correlation scores than when using GloVe embeddings in two cases: +0.181 on the Arabic task, +0.129 on the Spanish task, and comparable results (-0.005) on the English track. We stress that no re-training was performed on LESSLEX vectors on languages different from English, so that the improvement obtained in tracks 1 and 3 (ara-ara and spa-spa, respectively) is even more relevant. We interpret this achievement as stemming from the fact that LESSLEX vectors contain both conceptual and terminological descriptions: This seems also to explain the fact that the advantage obtained by using LESSLEX vectors with respect to GloVe is more sensible for languages where the translation and/or re-training are less effective, such as pairs involving either the Arabic or Turkish language. Also, we note that using contextual information (c-rank strategy) to govern the selection of senses ensures comparable results with the t-rank strategy across settings (with the exception of track 4b, where the drop in the correlation is very

prominent—one order of magnitude). Finally, it is interesting to observe that in dealing with crosslingual texts that involve arguably less-covered languages (i.e., in tracks 2 and 6, ara-eng and tur-eng), the c-rank strategy produced better results than the t-rank strategy.

To summarize the results on the STS task, by plugging LESSLEX embeddings into a state-of-the-art system such as HCTI we obtained results that either improve or are comparable to more computationally intensive approaches involving either MT or re-training, necessary to use GLoVe vectors in a multilingual and crosslingual setting. One distinguishing feature of our approach is that of hosting terminological and conceptual information in the same semantic space: Experimental evidence seems to confirm it as helpful in reducing the need for further processing, and beneficial to map different languages onto such unified semantic space.

4.4 General Discussion

Our experimentation has taken into account overall 11 languages, from different linguistic lineages, such as Arabic, coming from the Semitic phylum; Basque, a language isolate (reminiscent of the languages spoken in southwestern Europe before Latin); English and German, two West Germanic languages; Farsi, which as an Indo-Iranian language can be ascribed to the set of Indo-European languages; Spanish and Portuguese, which are Western Romance languages in the Iberian-Romance branch; French, from the Gallo-Romance branch of Western Romance languages; Italian, also from the Romance lineage; Russian, from the eastern branch of the Slavic family of languages; Turkish, in the group of Altaic languages, featured by phenomena such as vowel harmony and agglutination.

We utilized LESSLEX embeddings in order to cope with three tasks: (i) the traditional semantic similarity task, where we experimented on six different data sets (RG-65, WS-Sim-353, SimLex-999, SimVerbs-3500, SemEval-2017 (Task 2) and Goikoetxea-2018); (ii) the contextual semantic similarity task, where we experimented on two data sets, SCWS and WiC; (iii) the STS task, where the STS Benchmark and the SemEval-2017 (Task 1) data set were used for the experimentation.

In the first mentioned task (Section 4.1) our experiments show that in most cases LESSLEX results improve on those by all other competitors. As competitors all the principal embeddings were selected that allow to cope with multilingual tasks: ConceptNet Numberbatch, NASARI, JOINTCHYCB, SENSEEmbed, and NASARI2Vec.

Two different experimental conditions were considered (MSV and CbA, Table 11). Both views on results indicate that our approach outperforms the existing ones. To the best of our knowledge this is the most extensive experimentation ever performed on as many benchmarks, and including results for as many resources.

In dealing with the Contextual Similarity task (Section 4.2) we compared our results with those obtained by using NASARI2VEC, which also contains descriptions for both terms and nominal concepts in the same semantic space, and with results available in literature. The obtained figures show that despite not being tuned for this task, our approach improves on previous results on the SCWS data set. On the WiC data set, results obtained by experimenting with LESSLEX vectors overcome all those provided by directly comparable resources. Results obtained by state-of-the-art approaches (using contextualized sense embeddings) in this task are about 9% above those currently achieved through sense embeddings.

With regard to the third task on Semantic Text Similarity (Section 4.3), we used our embeddings by feeding them to a Convolutional Neural Network in place of GloVe

embeddings. The main outcome of this experiment is that although our results are comparable to those obtained by using GloVe for English tracks, they improve on the results obtained with GloVe in the crosslingual setting, even though these are specifically retrained on the considered tracks.

In general, handling sense-embeddings involves some further processing to select senses for input terms, while with word-embeddings one can typically benefit from the direct mapping term-vector. Hence, the strategy used to select senses is relevant when using LessLex embeddings. Also—though indirectly—subject to evaluation was the proposed similarity metrics of ranked-similarity; it basically relies on ranking sense vectors based on their distance from the terminological one. Ranked-similarity clearly outperforms the maximization of cosine similarity on LessLex embeddings. Besides, the contextual ranked-similarity (which was devised to deal with the contextual similarity task) was shown to perform well, by taking into account information from the context vector rather than from the terminological one. We defer to further work an exhaustive exploration of their underlying assumptions and the analytical description of differences in computing conceptual similarity between such variants of ranked similarity and existing metrics such as, for example, the Rank-Biased Overlap.

5. Conclusions

As illustrated in the discussion of results, the experimentation provides solid evidence that LessLex obtains competitive results with state-of-the-art word embeddings. In addition, LessLex provides conceptual grounding on BabelNet, one of the largest existing multilingual semantic networks. This enables the usage of LessLex vectors in conjunction with a broad web of senses that provides lexicalizations for 284 different languages, dealing with verbs, nouns, and adjectives, along with a rich set of semantic relations. The linking with BabelNet's sense inventory allows us to directly plug LessLex vectors into applications and to pair LessLex with resources that already adopt BabelNet synset identifiers as naming convention, such as DBPedia and Wikidata. The obtained results show that LessLex is a natural (vectorial) counterpart for BabelNet, favorably comparing to NASARI in all considered tasks, and also including verbs and adjectives that were left out of NASARI by construction.

LessLex adopts a unique semantic space for concepts and terms from different languages. The comparison with JointChyb shows that using a common conceptual level can be experimentally advantageous over handling specialized bilingual embeddings. Far from being an implementation feature, the adopted semantic space describes a cognitively plausible space, compatible with the cognitive mechanisms governing lexical access, which is in general featured by conceptual mediation (Marconi 1997). Further investigations are possible, stemming from the fact that senses and terms share the same multilingual semantic space: For example, we are allowed to compare and unveil meaning connections between terms across different languages. Such capabilities can be useful in characterizing subtle and elusive meaning shift *phenomena*, such as diachronic sense modeling (Hu, Li, and Liang 2019) and conceptual misalignment, which is a well-known issue, for example, in the context of automatic translation. This issue has been approached, for the translation of European laws, through the design of formal ontologies (Ajani et al. 2010).

Acquiring vector descriptions for concepts (as opposed to terms) may also be beneficial to investigate the conceptual abstractness/concreteness issue (Hill, Korhonen, and Bentz 2014; Mensa, Porporato, and Radicioni 2018; Colla et al. 2018), and its contribution to lexical competence (Paivio 1969; Marconi 1997). Characterizing concepts on

abstractness accounts is relevant for both scientific and applicative purposes. The investigation on abstract concepts has recently emerged as central in the multidisciplinary debate between grounded views of cognition versus modal (or symbolic) views of cognition (Bolognesi and Steen 2018). In the first hypothesis cognition might be embodied and grounded in perception and action (Gibbs Jr 2005): That is, accessing concepts would amount to retrieving and instantiating perceptual and motoric experience. Conversely, modal approaches to concepts are mostly in the realm of distributional semantic models: in this view the meaning of *rose* would result from "statistical computations from associations between *rose* and concepts like *flower*, *red*, *thorny*, and *love*" (Louwerse 2011, page 2). Also, accounting for conceptual abstractness may be beneficial in diverse NLP tasks, like WSD (Kwong 2008), the semantic processing of figurative uses of language (Turney et al. 2011; Neuman et al. 2013), automatic translation and simplification (Zhu, Bernhard, and Gurevych 2010), the processing of social tagging information (Benz et al. 2011), and many others, as well.

Finally, we mention the proposed similarity measure, ranked-similarity. Preliminary tests, not reported in this work, showed that ranked similarity mostly outperforms the maximization of cosine similarity, to such an extent that the results of LessLex vectors reported in the Evaluation Section were computed with ranked-similarity. Such novel measure originates from a simple intuition: In computing conceptual similarity, scanning and comparing each and every sense available in some fine-grained sense inventory may be unnecessary and confusing. Instead, we rank senses using their distance from the term; top-ranked senses are more relevant, so that the formula to compute ranked-similarity refines cosine similarity by adding a mechanism for filtering and clustering senses based on their salience.

In this work we have proposed LessLex vectors. Such vectors are built by rearranging distributional descriptions around senses, rather than terms. These have been tested on the word similarity task, on the contextual similarity task, and on the semantic text similarity task, providing good to outstanding results, on all data sets utilized. We have discussed the obtained results. Also importantly, we have outlined the relevance of LessLex vectors in the broader context of research in natural language with focus on senses and conceptual representation, mentioning that having co-located sense and term representations may be helpful to investigate some issues in an area at the intersection of general Artificial Intelligence, Cognitive Science, Cognitive Psychology, Knowledge Representation, and, of course, Computational Linguistics. In these settings distributed representation of senses may be used, either to enable further research or to solve specific tasks. In so doing, we feel that this work to some extent takes up a famous challenge, "to think about problems, architectures, cognitive science, and the details of human language, how it is learned, processed, and how it changes, rather than just chasing state-of-the-art numbers on a benchmark task" (Manning 2015, page 706).

Appendix A

A.1 Results on the Word Similarity Task, CbA Condition

In this section we illustrate the results obtained by testing on the semantic similarity task. However, different from the results reported in Section 4.1.3, in this case only the fraction of each data set covered by all considered resources was used for testing.

Table 17
Results on the subset of the multilingual and crosslingual RG-65 data set containing only word pairs covered by all considered resources. Reported figures express Pearson (*r*) and Spearman (ρ) correlations. In the first column we report the coverage for each translation of the data set actually used in the experimentation.

RG-65	LL-M		LLX		CNN		NAS		JCH		SSE		N2V	
	r	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ
[Word] eng [1.0]	.64	.59	**.91**	.86	**.91**	**.90**	.67	.67	.84	.86	.75	.81	.80	.75
[Sense] eng [1.0]	–	–	**.94**	**.91**	–	–	.81	.76	–	–	.72	.76	.78	.73
fas (N) [.69]	.78	.73	.86	.87	**.88**	**.89**	.71	.69	–	–	.72	.60	–	–
spa (N) [.98]	.82	.82	**.92**	**.93**	**.92**	**.93**	.91	.91	.80	.83	.82	.84	–	–
por-fas (N) [.81]	.73	.72	.91	**.90**	**.93**	.89	.79	.76	–	–	.76	.70	–	–
fra-por (N) [.97]	.83	.84	**.93**	**.89**	**.93**	**.89**	.76	.69	–	–	.81	.73	–	–
fra-fas (N) [.87]	.72	.72	.90	.88	**.93**	**.89**	.73	.69	–	–	.74	.68	–	–
fra-spa (N) [.99]	.81	.80	**.93**	**.91**	**.93**	.89	.85	.83	–	–	.88	.86	–	–
fra-deu (N) [.99]	.82	.86	**.91**	**.90**	.89	.88	.81	.78	–	–	.78	.76	–	–
spa-por (N) [.98]	.83	.83	**.93**	**.92**	**.93**	**.92**	.83	.81	–	–	.80	.79	–	–
spa-fas (N) [.82]	.71	.69	.92	**.92**	**.93**	.91	.83	.82	–	–	.78	.83	–	–
eng-por (N) [.99]	.74	.72	**.94**	**.90**	.92	**.90**	.79	.76	–	–	.80	.77	–	–
eng-fas (N) [.83]	.68	.61	.92	.89	**.93**	**.92**	.79	.74	–	–	.78	.74	–	–
eng-fra (N) [1.0]	.71	.70	**.94**	**.92**	.92	.91	.76	.73	–	–	.81	.75	–	–
eng-spa (N) [.99]	.73	.71	**.93**	**.93**	**.93**	.92	.85	.85	.84	.85	.80	.85	–	–
eng-deu (N) [.98]	.74	.72	**.92**	**.90**	.90	**.90**	.83	.81	–	–	.77	.80	–	–
deu-por (N) [.96]	.89	.86	**.93**	**.89**	.92	.88	.82	.78	–	–	.77	.74	–	–
deu-fas (N) [.81]	.76	.74	**.92**	**.91**	**.92**	.90	.88	.81	–	–	.82	.82	–	–
deu-spa (N) [.97]	.85	.86	**.92**	**.91**	.91	.90	.89	.86	–	–	.80	.81	–	–

Table 18
Results on the subset of the WS-Sim-353 dat aset containing only word pairs covered by all considered resources. Reported figures express Pearson (*r*) and Spearman (ρ) correlations. In the first column we report the coverage for each translation of the dat aset actually used in the experimentation.

WS-Sim-353	LL-M		LLX		CNN		NAS		JCH		SSE		N2V	
	r	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ
eng (N) [.97]	.67	.65	**.78**	**.79**	**.78**	**.79**	.60	.61	.75	.76	.69	.73	.71	.70
ita (N) [.92]	.68	.69	.74	**.77**	**.75**	**.77**	.66	.65	.69	.70	.65	.71	–	–
deu (N) [.88]	.77	.74	.83	.81	**.84**	**.83**	.70	.69	–	–	.65	.64	–	–
rus (N) [.83]	.75	.76	.77	.78	**.79**	**.79**	.66	.66	–	–	.63	.64	–	–

Table 19
Results on the subset of the SimVerbs-3500 data set containing only word pairs covered by all considered resources. Reported figures express Pearson (*r*) and Spearman (ρ) correlations. In the first column we report the coverage for each translation of the data set actually used in the experimentation.

SimVerbs-3500	LL-M		LLX		CNN		NAS		JCH		SSE		N2V	
	r	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ
eng (V)[1.0]	.58	.56	**.67**	**.66**	.62	.60	–	–	.56	.56	.45	.42	.31	.30

Table 20
Results on the subset of the multilingual SimLex-999 containing only word pairs covered by all considered resources. Reported figures express Pearson (r) and Spearman (ρ) correlations. In the first column we report the coverage for each translation of the data set actually used in the experimentation.

SimLex-999	LL-M		LLX		CNN		NAS		JCH		SSE		N2V	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
eng (N)[1.0]	.51	.52	**.69**	**.67**	.66	.63	.41	.39	.55	.53	.52	.49	.46	.44
eng (V) [1.0]	.62	.56	**.67**	**.65**	.61	.58	–	–	.51	.50	.54	.49	–	–
eng (A) [1.0]	**.84**	**.83**	.82	.79	.80	.78	–	–	.63	.62	.55	.51	–	–
eng (*) [1.0]	.57	.53	**.70**	**.69**	.67	.65	–	–	.55	.54	.53	.49	–	–
ita (N) [.96]	.50	.49	**.66**	**.64**	.64	.62	.48	.49	.48	.49	.56	.50	–	–
ita (V) [.96]	.58	.53	**.70**	**.63**	.69	.59	–	–	.57	.50	.56	.45	–	–
ita (A) [.95]	.68	.57	**.77**	**.70**	.73	.64	–	–	.40	.30	.61	.49	–	–
ita (*) [.96]	.49	.43	**.67**	**.63**	.65	.62	–	–	.48	.46	.55	.48	–	–
deu (N) [.94]	.58	.57	.66	.65	**.68**	**.66**	.46	.47	–	–	.48	.44	–	–
deu (V) [.73]	.56	.53	.63	**.60**	**.64**	.58	–	–	–	–	.51	.46	–	–
deu (A) [.67]	.74	.70	.76	.73	**.80**	**.75**	–	–	–	–	.50	.39	–	–
deu (*) [.86]	.59	.57	.66	.65	**.69**	**.67**	–	–	–	–	.47	.42	–	–
rus (N) [.86]	.45	.43	**.54**	**.51**	**.54**	.49	.23	.23	–	–	.26	.21	–	–
rus (V) [.20]	.60	.54	.58	.59	**.66**	**.60**	–	–	–	–	.42	.28	–	–
rus (A) [.06]	.92	.87	**.94**	**.91**	**.94**	.87	–	–	–	–	.62	.24	–	–
rus (*) [.63]	.46	.44	**.55**	**.51**	**.55**	.50	–	–	–	–	.27	.21	–	–

Table 21
Results on the subset of the SemEval 17 Task 2 data set containing only word pairs covered by all considered resources. Reported figures express Pearson (r) and Spearman (ρ) correlations. In the first column we report the coverage for each translation of the data set actually used in the experimentation.

SemEval-2017	LL-M		LLX		CNN		NAS		JCH		SSE		N2V	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
eng (N)[.66]	.70	.70	**.84**	**.86**	.83	.85	.57	.59	.75	.77	.71	.75	.73	.73
deu (N) [.73]	.78	.79	**.84**	.85	**.84**	**.86**	.68	.68	–	–	.67	.69	–	–
ita (N) [.61]	.73	.73	**.82**	**.84**	.80	.82	.75	.76	.76	.78	.71	.77	–	–
spa (N) [.62]	.77	.79	**.84**	**.86**	.81	.84	.70	.71	.78	.80	.73	.78	–	–
fas (N) [.34]	.69	.72	**.79**	**.82**	.75	.80	.58	.59	–	–	.65	.70	–	–
deu-spa (N) [.73]	.78	.80	**.84**	**.86**	.82	.84	.71	.72	–	–	.70	.74	–	–
deu-ita (N) [.74]	.77	.78	**.83**	**.85**	.82	.84	.72	.73	–	–	.69	.73	–	–
eng-deu (N) [.82]	.78	.79	**.85**	**.86**	.83	.85	.67	.68	–	–	.70	.72	–	–
eng-spa (N) [.63]	.74	.75	**.85**	**.87**	.83	.85	.65	.66	.75	.78	.72	.77	–	–
eng-ita (N) [.62]	.73	.74	**.85**	**.87**	.83	.85	.69	.70	.73	.75	.72	.77	–	–
spa-ita (N) [.61]	.75	.76	**.84**	**.86**	.81	.84	.74	.74	.70	.71	.72	.78	–	–
deu-fas (N) [.49]	.75	.78	**.84**	**.86**	.81	.85	.71	.72	–	–	.69	.74	–	–
spa-fas (N) [.49]	.72	.74	**.84**	**.86**	.80	.84	.70	.72	–	–	.70	.77	–	–
fas-ita (N) [.49]	.71	.72	**.81**	**.84**	.72	.82	.70	.72	–	–	.69	.75	–	–
eng-fas (N) [.54]	.70	.71	**.82**	**.85**	.79	.82	.65	.68	–	–	.70	.75	–	–

Table 22
Results on the subset of the Goikoetxea data set containing only word pairs covered by all considered resources. Reported figures express Pearson (*r*) and Spearman (ρ) correlations. In the first column we report the coverage for each translation of the data set actually used in the experimentation.

Goikoetxea	LL-M		LLX		CNN		NAS		JCH		SSE		N2V	
	r	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ	*r*	ρ
spa-eus (N)[.75]	.75	.71	.80	**.74**	**.81**	.73	.74	.73	.69	.66	.74	.70	–	–
eng-eus (N)[.77]	.75	.72	**.93**	**.91**	**.93**	.90	.91	.90	.87	.84	.84	.86	–	–
eng-spa (N)[.99]	.73	.71	**.93**	**.93**	**.93**	.92	.85	.85	.84	.85	.80	.85	–	–
eus-ita (N)[.72]	.62	.66	**.69**	**.73**	.67	.63	.57	.59	.58	.63	.53	.56	–	–
spa-ita (N)[.93]	.60	.65	**.67**	**.75**	.66	.74	.58	.59	.56	.61	.53	.59	–	–
spa-eus (N)[.73]	.67	.70	**.74**	**.79**	.71	.78	.66	.67	.70	.74	.60	.64	–	–
eng-ita (N)[.96]	.59	.64	**.70**	.76	**.70**	**.77**	.51	.52	.61	.66	.51	.58	–	–
eng-eus (N)[.75]	.64	.67	**.75**	**.80**	.74	**.80**	.58	.60	.72	.76	.58	.63	–	–
eng-spa (N)[.97]	.62	.66	**.72**	**.78**	.71	**.78**	.55	.56	.68	.74	.57	.64	–	–
eng-spa (N)[.97]	.50	.49	**.67**	**.65**	.64	.62	.52	.51	.56	.52	.55	.52	–	–
eng-spa (V)[.96]	.53	.49	**.62**	**.60**	.59	.57	–	–	.48	.46	.53	.49	–	–
eng-spa (A)[.80]	.76	**.77**	**.77**	**.77**	**.77**	**.77**	–	–	.59	.60	.56	.50	–	–
eng-spa (*)[.95]	.54	.52	**.67**	**.66**	.65	.64	–	–	.54	.52	.55	.51	–	–
eng-ita (N)[.97]	.53	.53	**.71**	**.69**	.68	.66	.46	.47	.53	.51	.55	.52	–	–
eng-ita (V)[.58[	.62	.55	**.71**	**.67**	.67	.60	–	–	.51	.45	.56	.46	–	–
eng-ita (A)[.80]	.79	.73	**.84**	**.78**	.78	.70	–	–	.41	.36	.61	.48	–	–
eng-ita (*)[.82]	.56	.53	**.72**	**.70**	.69	.67	–	–	.50	.48	.56	.50	–	–
spa-ita (N)[.96]	.53	.53	**.68**	**.67**	.66	.65	.47	.49	.48	.47	.56	.54	–	–
spa-ita (V)[.56]	.56	.52	**.65**	**.60**	.64	.58	–	–	.47	.42	.56	.49	–	–
spa-ita (A)[.78]	.73	.66	**.79**	**.73**	.76	.69	–	–	.43	.38	.63	.51	–	–
spa-ita (*)[.80]	.55	.53	**.68**	**.66**	.67	.65	–	–	.47	.45	.56	.51	–	–

Acknowledgments
We thank the anonymous reviewers for many useful comments and suggestions: their work helped to substantially improve this article. We are also grateful to Sergio Rabellino, Simone Donetti, and Claudio Mattutino from the Technical Staff of the Computer Science Department of the University of Turin for their precious support with the computing infrastructures. Finally, thanks are due to the Competence Centre for Scientific Computing (C3S) of the University of Turin (Aldinucci et al. 2017).

References
Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, Savannah, GA.

Agirre, Eneko, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and Wordnet-based approaches. In *Proceedings of NAACL*, NAACL '09, pages 19–27, Boulder, CO.

Ajani, Gianmaria, Guido Boella, Leonardo Lesmo, Alessandro Mazzei, Daniele P. Radicioni, and Piercarlo Rossi. 2010. Multilevel legal ontologies. *Lecture Notes in Computer Science*, 6036 LNAI:136–154.

Aldarmaki, Hanan and Mona Diab. 2019. Context-aware crosslingual mapping. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3906–3911, Minneapolis.

Aldarmaki, Hanan, Mahesh Mohan, and Mona Diab. 2018. Unsupervised word mapping using structural similarities in monolingual embeddings. *Transactions of the Association for Computational Linguistics*, 6:185–196.

Aldinucci, Marco, Stefano Bagnasco, Stefano Lusso, Paolo Pasteris, Sergio Rabellino, and Sara Vallero. 2017. OCCAM: A flexible, multi-purpose and extendable

HPC cluster. *Journal of Physics: Conference Series*, 898(8):082039–082047.
Andreas, Jacob and Dan Klein. 2014. How much do word embeddings encode about syntax? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 822–827, Baltimore, MD.
Ansell, Alan, Felipe Bravo-Marquez, and Bernhard Pfahringer. 2019. An ELMo-inspired approach to SemDeep-5's Word-in-Context task. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) 2019*, 10(2):62–66.
Arora, Sanjeev, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495.
Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. 2018. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019, New Orleans, LA.
Artetxe, Mikel and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot crosslingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.
Baker, Collin F. and Christiane Fellbaum. 2009. WordNet and FrameNet as complementary resources for annotation. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 125–129, Singapore.
Bansal, Mohit, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 809–815, Baltimore, MD.
Benz, Dominik, Christian Körner, Andreas Hotho, Gerd Stumme, and Markus Strohmaier. 2011. One tag to bind them all: Measuring term abstractness in social metadata. In *Proceedings of ESWC*, pages 360–374, Berlin.
Berant, Jonathan and Percy Liang. 2014. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1415–1425, Baltimore, MD.
Bojar,Ondrej, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. 2014. Findings of the 2014 Workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, MD.
Bolognesi, Marianna and Gerard Steen. 2018. Editors' introduction: Abstract concepts: Structure, processing, and modeling. *Topics in Cognitive Science*, 10(3):490–500.
Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon.
Budanitsky, Alexander and Graeme Hirst. 2006. Evaluating Wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.
Chandar, Sarath AP, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C. Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems*, pages 1853–1861, Montreal.
Camacho-Collados, Jose and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788.
Camacho-Collados, Jose, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. Semeval-2017 task 2: Multilingual and crosslingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26, Vancouver.
Camacho-Collados, José, Mohammad Taher Pilehvar, and Roberto Navigli. 2015a. A framework for the construction of monolingual and cross-lingual word similarity data sets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 1–7, Beijing.
Camacho-Collados, José, Mohammad Taher Pilehvar, and Roberto Navigli. 2015b. NASARI: A novel approach to a semantically-aware representation of items. In *Proceedings of NAACL*, pages 567–577, Denver, CO.
Camacho-Collados, José, Mohammad Taher Pilehvar, and Roberto Navigli. 2016.

NASARI: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

Cambria, Erik, Robert Speer, Catherine Havasi, and Amir Hussain. 2010. SenticNet: A publicly available semantic resource for opinion mining. In *2010 AAAI Fall Symposium*, pages 14–18, Arlington, VA.

Cer, Daniel, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver.

Chen, Tao, Ruifeng Xu, Yulan He, and Xuan Wang. 2015. Improving distributed representation of word sense via Wordnet gloss composition and context clustering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 15–20, Beijing.

Chen, Xinxiong, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035, Doha.

Chi, Ta Chung and Yun-Nung Chen. 2018. CLUSE: Crosslingual unsupervised sense embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 271–281, Brussels.

Chi, Ta Chung, Ching-Yen Shih, and Yun-Nung Chen. 2018. BCWS: Bilingual contextual word similarity. *arXiv preprint arXiv:1810.08951*.

Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha.

Chollet, François, and others. 2015. Keras: The python deep learning library, Astrophysics Source Code Library, 2018.

Colla, Davide, Enrico Mensa, Aureliano Porporato, and Daniele P. Radicioni. 2018. Conceptual abstractness: from nouns to verbs. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, pages 70–75, Turin.

Conneau, Alexis, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen.

Conneau, Alexis, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Coulmance, Jocelyn, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Trans-gram, fast crosslingual word-embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1109–1113, Copenhagen.

Davies, Mark. 2009. The 385+ million word corpus of contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2):159–190.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis.

Duong, Long, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1285–1295, Austin, TX.

Faruqui, Manaal and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg.

Faruqui, Manaal and Chris Dyer. 2015. Non-distributional word vector representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 464–469, Beijing.

Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Flekova, Lucie and Iryna Gurevych. 2016. Supersense embeddings: A unified model for supersense interpretation, prediction, and utilization. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2029–2041, Berlin.

Ganitkevitch, Juri, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of NAACL-HLT*, pages 758–764, Atlanta, GA.

Gerz, Daniela, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simverb-3500: A large-scale evaluation set of verb similarity. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2182, Austin, TX.

Gibbs Jr., Raymond W. 2005. *Embodiment and Cognitive Science*. Cambridge University Press.

Goikoetxea, Josu, Aitor Soroa, and Eneko Agirre. 2018. Bilingual embeddings with random walks over multilingual Wordnets. *Knowledge-Based Systems*, 150(C):218–230.

Gouws, Stephan, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast bilingual distributed representations without word alignments. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 748–756, Lille.

Guo, Mandy, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-hsuan Sung, Brian Strope, and others. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176, Brussels.

Harris, Zellig S. 1954. Distributional structure. *Word*, 10(2–3):146–162.

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, and others. 2018. Achieving human parity on automatic Chinese to English news translation. *arXiv preprint arXiv:1803.05567*.

Havasi, Catherine, Robert Speer, and Jason Alonso. 2007. ConceptNet: A lexical resource for common sense knowledge. *Recent Advances in Natural Language Processing V: Selected Papers from RANLP*, 309:269.

Hill, Felix, Anna Korhonen, and Christian Bentz. 2014. A quantitative empirical analysis of the abstract/concrete distinction. *Cognitive Science*, 38(1):162–177.

Hill, Felix, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Hisamoto, Sorami, Kevin Duh, and Yuji Matsumoto. 2013. An empirical investigation of word representations for parsing the web. In *Proceedings of ANLP*, pages 188–193, Nagoya.

Hliaoutakis, Angelos, Giannis Varelas, Epimenidis Voutsakis, Euripides G. M. Petrakis, and Evangelos Milios. 2006. Information retrieval by semantic similarity. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2(3):55–73.

Hu, Renfen, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 3899–3908, Florence.

Huang, Eric H., Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882, Jeju Island.

Iacobacci, Ignacio, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. SensEmbed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 95–105, Beijing.

Jimenez, Sergio, Claudia Becerra, Alexander Gelbukh, Av Juan Dios Bátiz, and Av Mendizábal. 2013. Softcardinality-core: Improving text overlap with distributional measures for semantic textual similarity. In *Proceedings of *SEM 2013*, volume 1, pages 194–201, Atlanta, GA.

Kenter, Tom and Maarten De Rijke. 2015. Short text similarity with word embeddings. In *Proceedings of the 24th ACM International Conference on Information*

and Knowledge Management, pages 1411–1420, New York, NY.

Kiros, Ryan, Yukun Zhu, Ruslan R. Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, pages 3294–3302.

Kočiskỳ, Tomáš, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning bilingual word representations by marginalizing alignments. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 224–229.

Kusner, Matt, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966, Lille.

Kwong, Olivia OY. 2008. A preliminary study on the impact of lexical concreteness on word sense disambiguation. In *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*, pages 235–244, Cebu City.

Lavie, Alon and Michael J Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23(2–3):105–115.

Le, Quoc and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196, Beijing.

Leviant, Ira and Roi Reichart. 2015a. Judgment language matters: Multilingual vector space models for judgment language aware lexical semantics. *CoRR, abs/1508.00106*.

Leviant, Ira and Roi Reichart. 2015b. Separated by an un-common language: Towards judgment language informed vector space modeling. *arXiv preprint arXiv:1508.00106*.

Lieto, Antonio, Enrico Mensa, and Daniele P. Radicioni. 2016a. A Resource-Driven Approach for Anchoring Linguistic Resources to Conceptual Spaces. In *XVth International Conference of the Italian Association for Artificial Intelligence*, pages 435–449, Genova.

Lieto, Antonio, Enrico Mensa, and Daniele P. Radicioni. 2016b. Taming sense sparsity: A common-sense approach. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*, pages 1–6, Napoli.

Lieto, Antonio, Daniele P. Radicioni, and Valentina Rho. 2015. A common-sense conceptual categorization system integrating heterogeneous proxytypes and the dual process of reasoning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 875–881, Buenos Aires.

Lieto, Antonio, Daniele P. Radicioni, and Valentina Rho. 2017. Dual PECCS: A cognitive system for conceptual representation and categorization. *Journal of Experimental & Theoretical Artificial Intelligence*, 29(2):433–452.

Logeswaran, Lajanugen and Honglak Lee. 2018. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*.

Loureiro, Daniel and Alipio Jorge. 2019. LIAAD at SemDeep-5 challenge: Word-in-Context (wic). In *Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5)*, pages 1–5.

Louwerse, Max M. 2011. Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 3(2):273–302.

Luong, Thang, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, Denver, CO.

Mancini, Massimiliano, Jose Camacho-Collados, Ignacio Iacobacci, and Roberto Navigli. 2017. Embedding words and senses together via joint knowledge-enhanced training. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 100–111, Vancouver.

Manning, Christopher D. 2015. Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707.

Marconi, Diego. 1997. *Lexical Competence*. MIT Press.

Mensa, Enrico, Aureliano Porporato, and Daniele P. Radicioni. 2018. Annotating concept abstractness by common-sense knowledge. In Ghidini, Chiara, Bernardo Magnini, Andrea Passerini, and Paolo Traverso, editors. *AI*IA 2018 – Advances in Artificial Intelligence*, pages 415–428, Trento.

Mensa, Enrico, Daniele P. Radicioni, and Antonio Lieto. 2017. MERALI at

SemEval-2017 Task 2 Subtask 1: A cognitively inspired approach. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 236–240, Vancouver.

Mensa, Enrico, Daniele P. Radicioni, and Antonio Lieto. 2018. COVER: A linguistic resource combining common sense and lexicographic information. *Language Resources and Evaluation*, 52(4):921–948.

Mikolov, Tomas, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Miller, George A. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Miller, George A. and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Minsky, Marvin. 1975. A framework for representing knowledge. In Winston, P., editor, *The Psychology of Computer Vision*, McGraw-Hill, New York, pages 211–277.

Mohammad, Saif M. and Graeme Hirst. 2012. Distributional measures of semantic distance: A survey. *arXiv preprint arXiv:1203.1858*.

Moro, Andrea, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: A unified approach. *Transactions of the Association for Computational Linguistics*, 2231–244.

Mrkšić, Nikola, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gasic, Lina M. Rojas Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148, San Diego, CA.

Mu, Jiaqi, Suma Bhat, and Pramod Viswanath. 2017. Geometry of polysemy. In *5th International Conference on Learning Representations, ICLR 2017, Conference Track Proceedings*, Toulon.

Navigli, Roberto. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 105–112, Sydney.

Navigli, Roberto and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala.

Navigli, Roberto and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*. 193:217–250.

Neelakantan, Arvind, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069, Doha.

Nelson, Douglas L., Cathy L. McEvoy, and Thomas A. Schreiber. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

Neuman, Yair, Dan Assaf, Yohai Cohen, Mark Last, Shlomo Argamon, Newton Howard, Ophir Frieder, et al. 2013. Metaphor identification in large texts corpora. *PLOS ONE*, 8(4):1–9.

Paivio, Allan. 1969. Mental imagery in associative learning and memory. *Psychological Review*, 76(3):241.

Palmer, Martha, Olga Babko-Malaya, and Hoa Trang Dang. 2004. Different sense granularities for different applications. In *Proceedings of the 2nd International Workshop on Scalable Natural Language Understanding (ScaNaLU 2004) at HLT-NAACL 2004*, pages 49–56, Boston, MA.

Pedersen, Ted, Satanjeev Banerjee, and Siddharth Patwardhan. 2005. Maximizing semantic relatedness to perform word sense disambiguation. *University of Minnesota Supercomputing Institute Research Report UMSI*, 25:2005.

Pelevina, Maria, Nikolay Arefiev, Chris Biemann, and Alexander Panchenko. 2016. Making sense of word embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 174–183, Berlin.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), volume 14, pages 1532–1543.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237, New Orleans, LA.

Pilehvar, Mohammad Taher and Jose Camacho-Collados. 2019. WiC: The Word-in-Context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, MN.

Pilehvar, Mohammad Taher and Nigel Collier. 2016. De-conflated semantic representations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1680–1690, Austin, TX.

Pilehvar, Mohammad Taher and Roberto Navigli. 2015. From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artificial Intelligence*, 228:95–128.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Reisinger, Joseph and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117, Los Angeles, CA.

Resnik, Philip. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th IJCAI*, pages 448–453, Montréal.

Rosch, Eleanor. 1975. Cognitive Representations of Semantic Categories. *Journal of Experimental Psychology: General*, 104(3):192–233.

Rubenstein, Herbert and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Ruder, Sebastian, Ivan Vulić, and Anders Søgaard. 2019. A survey of crosslingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Schuster, Tal, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Crosslingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1599–1613, Minneapolis, MN.

Schwartz, Hansen A. and Fernando Gomez. 2011. Evaluating semantic metrics on tasks of concept similarity. In *Proceedings of the International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pages 299–304, Palm Beach, FL.

Schwenk, Holger and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. *ACL 2017*, pages 157–167, Vancouver.

Shao, Yang. 2017. HCTI at SemEval-2017 Task 1: Use convolutional neural network to evaluate semantic textual similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 130–133, Vancouver.

Soler, Aina Garí, Marianna Apidianaki, and Alexandre Allauzen. 2019. LIMSI-MULTISEM at the IJCAI SemDeep-5 WiC Challenge: Context representations for word usage similarity estimation. In *Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5)*, pages 6–11, Macaur.

Speer, Robert and Joshua Chin. 2016. An ensemble method to produce high-quality word embeddings. *arXiv preprint arXiv:1604.01692*.

Speer, Robert, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, pages 4444–4451.

Speer, Robert and Catherine Havasi. 2012. Representing general relational knowledge in ConceptNet 5. In *LREC*, pages 3679–3686.

Speer, Robyn and Joanna Lowry-Duda. 2017. ConceptNet at SemEval-2017 Task 2: Extending word embeddings with multilingual relational knowledge. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 85–89, Istanbul.

Tang, Duyu, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 1555–1565, Baltimore, MD.

Turney, Peter D., Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh.

Tversky, Amos. 1977. Features of similarity. *Psychological Review*, 84(4):327.

Vulić, Ivan and Anna Korhonen. 2016. On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 247–257, Berlin.

Vulić, Ivan and Marie-Francine Moens. 2015. Monolingual and crosslingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 363–372, Santiago.

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3261–3275, Santiago.

Webber, William, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):20.

Zhu, Zhemin, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International conference on Computational linguistics*, pages 1353–1361, Uppsala.