A Methodology for Large-Scale, Disambiguated and Unbiased Lexical Knowledge Acquisition Based on Multilingual Word Alignment

Francesca Grasso, Luigi Di Caro
University of Turin, Department of Computer Science
{fr.grasso,luigi.dicaro}@unito.it

Abstract

In order to be concretely effective, many NLP applications require the availability of lexical resources providing varied, broadly shared, and language-unbounded lexical information. However, state-of-the-art knowledge models rarely adopt such a comprehensive and cross-lingual approach to semantics. In this paper, we propose a novel automatable methodology for knowledge modeling based on a multilingual word alignment mechanism that enhances the encoding of unbiased and naturally disambiguated lexical knowledge. Results from a simple implementation of the proposal show relevant outcomes that are not found in other resources.

1 Introduction

Lexical resources constitute a key instrument for many NLP tasks such as Word Sense Disambiguation and Machine Translation. However, their potential may vary widely depending on the nature of the lexical-semantic knowledge they encode, as well as on how the linguistic data are stored and linked within the network (Zock and Biemann, 2020). The resources that are presently available, such as WordNet (Miller, 1995), typically encode lexical-semantic knowledge mainly in terms of word senses, defined by textual (i.e. dictionary) definitions, and lexical entries are linked and put in context through lexical-semantic relations. These relations, being only of a paradigmatic nature, are characterized by a sharing of the same defining properties between the words and a requirement that the words be of the same syntactic class (Morris and Hirst, 2004). Typically related words are therefore not represented due to the absence of syntagmatic links. Additionally, word senses suffer from a lack of explicit common-sense knowledge and context-dependent information. Finally, the well-known fine granularity of word senses in WordNet (Palmer et al., 2007) is due to the lack of a meaning encoding system capable of representing concepts in a flexible way. Other kinds of resources such as FrameNet (Baker et al., 1998) and ConceptNet (Speer et al., 2017) present the same issue, while returning different types and degrees of structural semantic information and disambiguation capabilities.

In this contribution, we provide a novel methodology for the retrieval and representation of unbiased and naturally disambiguated lexical information that relies on a multilingual word alignment mechanism. In particular, we exploit textual resources in different languages[1] in order to acquire and align varied lexical-semantic material of the form $<$*target-concept*, $\{$*related words*$\}^k>$ that are common and shared by all the k languages involved. As we demonstrate through a simple implementation, our method allows to create new lexical-semantic relations between words that are not always available in other resources, as well as to perform an automatic word sense disambiguation process. This system therefore enhances the encoding of prototypical semantic information of concepts that is also likely to be free from strong cultural-linguistic and lexicographic biases.

The benefits provided by our novel multilingual word alignment mechanism are thus fourfold: *(i)* a linguistic and lexicographic de-biasing of lexical knowledge; *(ii)* naturally-disambiguated aligned lexical entries; *(iii)* the discovery of novel lexical-semantic relations; and *(iv)* the representation of prototypical semantic information of concepts in different languages.

[1] In this work, we start with the combination of three languages: English, German and Italian.

2 Background and Related Work

2.1 Bias Types

Due to its complex and fluid nature, lexical semantics needs to undergo a process of abstraction and simplification in order to be encoded into a formal model. As a result, lexical knowledge provided by lexical resources - especially when monolingual - will inherently carry different types of biases. In particular, *i)* linguistic and *ii)* lexicographic biases affect the encoding, consumption, and exploitation of lexical knowledge in downstream tasks.

Linguistic bias Lexical information encoded in a language's lexicon, as well as the potential contexts in which a given lexeme can occur, inevitably reflect the socio-cultural background of the speakers of that language. Lexical resources used for the compilation of lexical knowledge are often conceived as monolingual, therefore they mostly return culture-bounded semantic information which does not account for more shared knowledge.

Lexicographic bias The nuclear components extracted from textual definitions can be different depending on the resource used, even within a single language (Kiefer, 1988). For example, the definition of "*cow*" reported by the Oxford Dictionary is "*a large animal kept on farms to produce milk or beef*" while the Merriam-Webster Dictionary reports "*the mature female of cattle*". Both endogenous and exogenous properties can be subjectively reported (Woods, 1975), such as the term "*large*" and the milk production respectively.

2.2 Related Work

On one side, lexicons are built on top of synsets[2] and contextualize meanings (or senses) mainly in terms of paradigmatic relations. WordNet (Miller, 1995) and BabelNet (Navigli and Ponzetto, 2010) can be seen as the cornerstone and the summit in that respect. However, if on the one hand WordNet's dense network of taxonomic relationships allows a high degree of systematization, on the other hand, a key unsolved issue with "*wordnets*" is the fine granularity of their inventories. Note that multilingualism in BabelNet is provided as an indexing service rather than as an alignment and unbiasing systematization method.

Extensions of these resources also include Common-Sense Knowledge (CSK), which refers to some (to a certain extent) widely-accepted and shared information. CSK describes the kind of general knowledge material that humans use to define, differentiate and reason about the conceptualizations they have in mind (Ruggeri et al., 2019). ConceptNet (Speer et al., 2017) is one of the largest CSK resources, collecting and automatically integrating data starting from the original MIT Open Mind Common Sense project[3]. However, terms in ConceptNet are not disambiguated. Property norms (McRae et al., 2005; Devereux et al., 2014) represent a similar kind of resource, which is more focused on the cognitive and perception-based aspects of word meaning. Norms, in contrast with ConceptNet, are based on semantic features empirically-constructed via questionnaires producing lexical (often ambiguous) labels associated with target concepts, without any systematic methodology of knowledge collection and encoding.

Another widespread modeling approach is based on vector space models of lexical knowledge. Vectors are automatically learnt from large corpora utilizing a wide range of statistical techniques, all centered on Harris' distributional assumption (Harris, 1954), i.e. words that occur in the same contexts tend to have similar meanings. Well-known models include word embeddings (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2016), sense embeddings (Huang et al., 2012; Iacobacci et al., 2015; Kumar et al., 2019), and contextualized embeddings (Scarlini et al., 2020). However, the relations holding between vector representations are not typed, nor are they organized systematically.

Among the several other modeling strategies proposed, lexicographic-centered resources have been focused on the contextualization of lexical items within syntactic structures, e.g. Corpus Pattern Analysis (CPA) (Hanks, 2004), situation frames such as FrameNet (Fillmore, 1977; Baker et al., 1998) and conceptual frames (Moerdijk et al., 2008; Leone et al., 2020). Words are not taken in isolation and the meaning they are attributed is connected to prototypical patterns or typed slots. However, these theories and methods for building semantic resources remain linked to the lexical basis and do not manage the mentioned biases.

[2] Words considered as synonyms in specific contexts.

[3] https://www.media.mit.edu/projects/open-mind-common-sense/overview/

3 The Multilingual Word Alignment

As is known, a single word form can be associated with more than one related sense, causing what is referred to as semantic ambiguity, or polysemy. This phenomenon, however, manifests itself differently across languages, since each language encodes meaning into words in its own particular way. We can therefore assume that, while a given polysemous word may be ambiguous in a certain context, a semantically corresponding word in another language will possibly not. Based on this assumption, it is possible to exploit this cross-language property to disambiguate a given word using its semantic equivalent in another language when they both occur in the same context. Such disambiguation process can take place because the two words feature different semantic - specifically, polysemous - behaviours. Accordingly, we developed a knowledge acquisition methodology that features the power of word sense disambiguation, relying on a multilingual $<$*target-concept*, $\{$*related words*$\}^k>$ alignment mechanism.

After providing a brief illustration of the languages we have selected for this first trial, we describe more in detail the methodology by using a basic example. Afterwards, a simple implementation of the proposed mechanism is presented.

3.1 Languages Involved

Among the benefits provided by the multilingual word alignment methodology we propose, one is that it prevents the represented lexical information from containing strong cultural-linguistic biases. This objective is pursued through the use of three different languages, reflecting in turn three diverse backgrounds. For this first trial we involved English, German and Italian. These languages were chosen primarily because we are proficient in them, therefore we are able to exert control over the data of our trial, as well as to interpret the results properly. Concurrently, given the nature of the methodology, it was necessary to select a set of languages with a certain degree of similarity in terms of shared lexical-semantic material. Indeed, the alignment mechanism can work and be effective as long as the lexical-semantic systems of the languages involved reflect a somewhat similar cultural-linguistic background. For example, we might expect languages to agree on the meanings of "carp", "cottage" and "sled" as long as speakers of these languages have comparable exposure to the relevant data. We would not expect a language spoken in a place without carps to have a word corresponding to "carp". The purpose of this project is not to forcibly identify universally valid semantic relationships, rather to not report biased information deriving from the use of data coming from a single linguistic context. For this reason, in our case the choice fell on European languages [4] (two Germanic languages and a Romance one).

wool	**Wolle**	**lana**
sheep	*Schal*	*cotone*
cotton	*spinnen*	*Biella*
synthetic	*Baumwolle*	*sintetica*
spin	*Rudolf*	*sciarpa*
scarf	*synthetisch*	*pecora*
mitten	*Schafe*	*filare*

Table 1: Unordered lists of single-language related words for $<$*wool* (EN), *Wolle* (DE), *lana* (IT)$>$.

3.2 Method

We now describe in detail the alignment mechanism through a basic example. Consider the following word forms: *wool* (EN); *Wolle* (DE); *lana* (IT), expressing a single target concept[5].

For each of the three lexical forms we collect a set of related words in terms of paradigmatic (e.g. synonyms) and syntagmatic (e.g. co-occurrences) relations. The target-related words can possibly be modifiers, verbs, or substantives. We thus obtain three different lists of words, one for each of the languages involved. The retrieved terms in the lists are still potentially ambiguous, since they refer to a lexical form rather than to a contextually defined concept. Table 1 provides a small excerpt of such unordered lists of related words.

The lexical data in the lists are subsequently compared and filtered in order to select only the semantic items that occur in all the lists, i.e., those shared by the three languages[6], in the reported example. The resulting words are thus aligned with their semantic counterparts, generating a set of aligned triplets, as shown in Table 2.

This multilingual word alignment provides, as a consequence, an automatic Word Sense Disambiguation system. Once the triplets are formed, their members will be indeed associated with a

[4] By "European" we refer to the European linguistic area.

[5] An absolute monosemy is, of course, realistically unreachable.

[6] This implies the presence of a translation step.

wool		Wolle		lana
sheep	↔	*Schafe*	↔	*pecora*
cotton	↔	*Baumwolle*	↔	*cotone*
synthetic	↔	*syntetisch*	↔	*sintetica*
spin	↔	*spinnen*	↔	*filare*
scarf	↔	*Schal*	↔	*sciarpa*

Table 2: Examples of aligned concept-related words for <*wool* (EN), *Wolle* (DE), *lana* (IT)>.

likely unique sense, i.e. the one coming from the intersection of all possible language-specific senses related to the three words. In other terms, the target-related words, once aligned, naturally identify (and provide) a common semantic context. As a consequence, potentially polysemous words are disambiguated through such context, without any support from sense repositories. For example, the context-consistent sense of the verb *to spin* (EN), which is a highly polysemous word in English, can be identified by selecting the only sense that is also shared by the other two aligned words, i.e. "*turn fibres into thread*". In fact, neither *spinnen* (DE) nor *filare* (IT) can possibly mean e.g. "rotate".

This mechanism generates a twofold effect: besides performing word sense disambiguation, it also provides lexical knowledge in the form of (paradigmatic and syntagmatic) lexical-semantic relations between words that is also language-unbounded. In the first place, the uncontrolled character of the data retrieval and alignment process offers the generation of novel lexical-semantic relations that are likely not available in other structured resources. Additionally, since the resulting set of words related to the target can be only the one shared by multiple languages, the lexical knowledge it encodes does not reflect a single cultural/linguistic background, rather a common and shared one. For example, in Table 1 the presence of the word "*Biella*" among the list of words related to "*lana*", probably refers to the fact that the Italian city Biella is (locally) famous for its wool, therefore the two words may co-occur frequently. Similarly, if we consider the alignment <*cat* (EN), *Katze* (DE), *gatto* (IT)>, a lexeme related to the English word form would be "*rain*", due to the well-known idiom "*it's raining cats and dogs*". However, neither "*Biella*" nor corresponding words for "*rain*" can possibly result in the lists of related words of the respective other languages, being language-specific items within those contexts. Therefore, the lexical information provided by the alignment mechanism will be free from strong cultural-linguistic biases. Finally, as illustrated in the next section, by exploiting multiple and differently built resources, we are able to reduce arbitrariness and lexicographic biases within the lexical knowledge represented.

4 Implementation

In this section we describe details and results of a simple implementation of the proposed alignment mechanism for the acquisition of disambiguated and unbiased lexical information. In particular, the system is composed of two main modules: a context generation and an alignment procedure. We finally report the results of an evaluation to highlight mainly *(i)* the autonomous disambiguation power of the approach, *(ii)* the quality of the alignments and their unbiased and syntagmatic nature, and *(iii)* the amount of unveiled lexical-semantic relations not covered by existing state-of-the-art resources such as BabelNet.

POS	scale	bilancia	Waage
noun	accuracy	precisione	Genauigkeit
noun	balance	equilibrio	Balance
noun	bulk	massa	Masse
noun	control	controllo	Kontrolle
noun	device	dispositivo	Gerät
noun	figure	cifra	Zahl
adj	accurate	preciso	genau
adj	smart	intelligente	intelligent
verb	indicate	indicare	zeigen
verb	set	regolare	einstellen

Table 3: 10 automatic alignments (out of 74) for the target concept <*scale* (EN), *bilancia* (IT), *Waage* (DE)> (BabelNet synset:00069470n).

4.1 Context for Multilingual Alignment

To retrieve the concept-related words for the multilingual alignment we made use of two textual resources: Sketch Engine (Kilgarriff et al., 2014) and the Leipzig Corpora Collection (Quasthoff et al., 2014). Through the former, we searched for related words with its tool named "Word Sketch" on the TenTen Corpus Family[7]. In particular, we were able to automatically collect words appearing in the following grammatical relations: "*mod-*

[7] https://www.sketchengine.eu/documentation/tenten-corpora

	00008050n	00069470n	00069470n	00062766n	00008364n	00008363n	
(en)	*libra*	*scale*	*plane*	*plane*	*bank*	*bank*	
(it)	*bilancia*	*bilancia*	*aereo*	*piano*	*banca*	*riva*	
(de)	*Waage*	*Waage*	*Flugzeug*	*Ebene*	*Bank*	*Ufer*	
triplets	**26**	**74**	**272**	**151**	**349**	**80**	
novel(en)	88,46%	87,84%	88,97%	89,40%	87,68%	91,25%	**88,9%**
novel(it)	76,92%	66,22%	75,74%	73,51%	75,64%	68,75%	**72,8%**
novel(de)	88,46%	74,32%	87,87%	84,11%	81,66%	76,25%	**82,1%**

Table 4: Alignments for six ambiguous concepts and percentage of unveiled *novel* relations in each language with respect to the BabelNet database. Some examples of triplets for the concept *scale-bilancia-Waage* (bn:*00069470n*) are shown in Table 3.

ifiers of w", "*adj. predicates of w*", "*verbs with w as subject*" and "*verbs with w as object*". The retrieved concept-related words are then lemmatized and marked with the suitable POS tags. Finally, we utilized the Leipzig Corpora Collection portal for searching additional context words in terms of left and right (POS-tagged) co-occurrences.

4.2 Multilingual Alignment

The Google Translate API was used for finding translations of related words in the three languages[8]. In particular, given a certain term t^{L1} in a language $L1$, we opted for retrieving all its possible translations into the other two languages ($L2$, $L3$). We then tried to match each translated item with the previously-retrieved sets of related words in $L2$, $L3$. Whenever the $[t^{L1} \leftrightarrow t^{L2}]$; $[t^{L1} \leftrightarrow t^{L3}]$ match succeeded, we finally checked any possible $[t^{L2} \leftrightarrow t^{L3}]$ match. If a $[t^{L1} \leftrightarrow t^{L2} \leftrightarrow t^{L3}]$ semantic equivalence occurs, then the alignment can take place. Table 3 shows an excerpt of automatic alignments for the concept *scale* (bn:*00069470n*).

4.3 Evaluation

Our aim is not to overcome state-of-the-art resources but rather to incorporate new and unbiased semantic relations from a novel multilingual alignment mechanism. In particular, we wanted to verify to what extent our knowledge acquisition method is able to unveil lexical relations yet uncovered by a state-of-the-art resource (BabelNet).

Thus, we first generated sets of related words from BabelNet in order to compare them with those produced and aligned by our (automatized) methodology. In particular, through the BabelNet API, we obtained the English, Italian, and German lexicalizations of the synsets connected to it, together with the words included in their glosses[9].

As test cases, we randomly picked 500 concepts constituting polysemous words in at least one of the three languages, obtaining non-empty alignments for 456 of them. In Table 4 we report the results of the alignment on six concepts.

Despite its limitations, our first implementation of the proposed methodology was able to discover a total of 76,152 multilingual alignments over the 456 concepts, with (on average) more than 80% novel semantic relations with respect to what is currently encoded in BabelNet across the three languages. Still, the extracted data represent mostly unbiased and disambiguated knowledge, leading towards the construction of a new large-scale and multilingual prototypical lexical database.

5 Conclusions and Future Work

In this paper we proposed an original methodology for acquiring and encoding lexical knowledge through a novel yet simple mechanism of multilingual alignment. The aim was to represent varied, disambiguated, and language-unbounded lexical knowledge by minimizing strong linguistic and lexicographic biases. A simple implementation and experimentation on 456 concepts carried to unveil around 76K aligned lexical-semantic features, of which more than 80% resulted new when compared with a current state-of-the-art resource such as BabelNet. Future directions include the use of more languages and large-scale runs over thousands of main concepts (Bentivogli et al., 2004; Di Caro and Ruggeri, 2019; Camacho-Collados and Navigli, 2017).

[8]No surrounding syntactic context for the words to align was available for more advanced Machine Translation.

[9]We used the SpaCy library to analyze, extract and lemmatize the text - `https://spacy.io`.

References

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90.

Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. 2004. Revising the wordnet domains hierarchy: semantics, coverage and balancing. In *Proceedings of the workshop on multilingual linguistic resources*, pages 94–101.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Jose Camacho-Collados and Roberto Navigli. 2017. Babeldomains: Large-scale domain labeling of lexical resources. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 223–228.

Barry J Devereux, Lorraine K Tyler, Jeroen Geertzen, and Billi Randall. 2014. The cslb concept property norms. *Behavior research methods*, 46(4):1119–1127.

Luigi Di Caro and Alice Ruggeri. 2019. Unveiling middle-level concepts through frequency trajectories and peaks analysis. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 1035–1042.

Charles J Fillmore. 1977. Scenes-and-frames semantics. *Linguistic structures processing*, 59:55–88.

Patrick Hanks. 2004. Corpus pattern analysis. In *Euralex Proceedings*, volume 1, pages 87–98.

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

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proc. of ACL*, pages 873–882.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. SensEmbed: learning sense embeddings for word and relational similarity. In *Proceedings of ACL*, pages 95–105.

Ferenc Kiefer. 1988. Linguistic, conceptual and encyclopedic knowledge: Some implications for lexicography. In T. Magay and J. Zigány, editors, *Proceedings of the 3rd EURALEX International Congress*, pages 1–10, Budapest, Hungary, sep. Akadémiai Kiadó.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: Ten years on. *The Lexicography*, 1(1):7–36.

Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. 2019. Zero-shot word sense disambiguation using sense definition embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681.

Valentina Leone, Giovanni Siragusa, Luigi Di Caro, and Roberto Navigli. 2020. Building semantic grams of human knowledge. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2991–3000.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behav. r. m.*, 37(4):547–559.

Tomas Mikolov, 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.

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

Fons Moerdijk, Carole Tiberius, and Jan Niestadt. 2008. Accessing the anw dictionary. In *Proc. of the workshop on Cognitive Aspects of the Lexicon*, pages 18–24.

Jane Morris and Graeme Hirst. 2004. Non-classical lexical semantic relations. In *Proceedings of the Computational Lexical Semantics Workshop at HLT-NAACL 2004*, pages 46–51, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proc. of ACL*, pages 216–225. Association for Computational Linguistics.

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Nat.Lan.Eng.*, 13(02):137–163.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–43.

Uwe Quasthoff, Dirk Goldhahn, and Thomas Eckart. 2014. Building large resources for text mining: The leipzig corpora collection. In *Text Mining*, pages 3–24. Springer.

Alice Ruggeri, Luigi Di Caro, and Guido Boella. 2019. The role of common-sense knowledge in assessing semantic association. *Journal on Data Semantics*, 8(1):39–56.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. In *Proceedings of the 34th Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence.

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

William A Woods. 1975. What's in a link: Foundations for semantic networks. In *Representation and understanding*, pages 35–82. Elsevier.

Michael Zock and Chris Biemann. 2020. Comparison of different lexical resources with respect to the tip-of-the-tongue problem. *Journal of Cognitive Science*, 21(2):193–252.