MultiAligNet: Cross-lingual Knowledge Bridges Between Words and Senses

MultiAligNet: Cross-lingual Knowledge Bridges between Words and Senses

Francesca Grasso[1⋆][0000-0001-8473-9491], Vladimiro Lovera Rulfi[2], and Luigi Di Caro[1][0000-0002-7570-637X]

[1] University of Turin, Turin, Italy - {fr.grasso@unito.it, luigi.dicaro@unito.it}
[2] University of Bologna, Bologna, Italy - {vladimiro.lovera@unibo.it}

Abstract. Numerous NLP applications rely on the accessibility to multilingual, diversified, context-sensitive, and broadly shared lexical semantic information. Standard lexical resources tend to first encode monolithic language-bounded senses which are eventually translated and linked across repositories and languages. In this paper, we propose a novel approach for the representation of lexical-semantic knowledge in - and shared from the origin by - multiple languages, based on the idea of k-Multilingual Concept (MC^k). MC^ks consist of multilingual alignments of semantically equivalent words in k different languages, that are generated through a defined linguistic context and linked via empirically determined semantic relations without the use of any sense disambiguation process. The MC^k model allows to uncover novel layers of lexical knowledge in the form of multifaceted conceptual links between naturally disambiguated sets of words. We first present the conceptualization of the MC^ks, along with the word alignment methodology that generates them. Secondly, we describe a large-scale automatic acquisition of MC^ks in English, Italian and German based on the exploitation of corpora. Finally, we introduce *MultiAlignNet*, an original lexical resource built using the data gathered from the extraction task. Results from both qualitative and quantitative assessments on the generated knowledge demonstrate both the quality and the novelty of the proposed model.

Keywords: Lexical Semantics · Multilingual alignments

1 Introduction

The exploitation of lexical resources constitutes a key issue for several Natural Language Processing tasks and applications. Many existing resources, such as WordNet [30], usually encode language-bounded lexical knowledge in the form of *word senses*, i.e., dictionary-oriented definitions of lexical entries which are linked and put in context through lexical-semantic relations. These relations, being only of a paradigmatic nature, are characterized by a sharing of similar defining properties between the words and a requirement that the items belong to the same syntactic category [32]. The fine-grained structure of such resources

⋆ Corresponding Author.

and the lack of syntagmatic associations, while allowing a high systematization of the linguistic data, determines an artificial abstraction that does not always reflect empirical reality. This is mainly due to the lack of a meaning encoding system capable of representing concepts in a flexible way [35].

Word Sense Disambiguation (WSD) is the task of determining the context-consistent meaning of a word from among all its possible senses by drawing from a sense repository [33]. Sense repositories may vary in terms of generality (from top-level and general purposes up to domain-specific ones) and completeness. WordNet is currently one of the most commonly adopted, with counterparts in other languages [5] and links with other resources, e.g. BabelNet [34]. While many works focused on raising the state-of-the-art performance, the improvement still stops at 81% of F-score when using WordNet as sense inventory [26, 3]. This is due to the difficulty to perform disambiguation, which constitutes one of the more complex and elusive processes of the semantic landscape even in human-to-human dialogues [13, 37]. Current state-of-the-art approaches are mainly devoted to create or link repositories rather than clustering existing senses. In this paper we propose a different approach, providing a natively cross-lingual view of the problem.

As is known, lexical ambiguity is a natural property of semantic systems which, however, mutates from language to language. Therefore, it may decrease when putting lexical items in reciprocal relation, i.e., when aligned. While a given language may provide only a single disambiguation context for a word, the use of parallel languages may indeed help further restrict word sense variability [21]. For example, the concept of "*discharge from an office or position*" may be encoded into the English verb form "*to fire*" which is however highly ambiguous, counting twelve different verbal senses in WordNet. The same concept is expressed by another polysemous term in Italian, i.e. "*licenziare*". However, the words *fire - licenziare* when associated with each other represent a bilingual encoding of that single concept which naturally avoids ambiguity, given that there are no other meanings that the two words may share. Thus, translations of a target word into one or more languages provide it a disambiguation context and may serve as sense labels [27]. Many works [8, 10, 1, 27, 12], have already shown the advantages of multilingual word alignments to perform Word Sense Disambiguation, although dwelling on the exploitation of either parallel corpora or multilingual wordnets, i.e, on already existing and pre-determined cross-lingual lexical material. In this work, we propose to leverage this property of languages for a broader purpose.

First, we propose a novel lexical-semantic encoding model bridging between words and senses called *k-Multilingual Concept* (MC^k), based on the abovementioned cross-lingual alignment in k different languages. As a second contribution, we present a large-scale automatic acquisition of MC^ks from several corpora in three languages (English, Italian, and German). This model enables the encoding of varied layers of lexical knowledge, in terms of both syntagmatic and paradigmatic relations, providing networks of diversified conceptual links between words in - and shared by - different languages. Through the proposed method we extracted a total of $21,514$ trilingual alignments belonging to three different types of Part-of-Speech tags (nouns, modifiers and verbs) for more than 1,047 input WordNet synsets. As final contribution, we publicly release a resource, called *MultiAligNet*, in two different versions, i.e. in *i)* vectorial and *ii)* graph-based forms. Finally, we evaluate the resource through both qualitative and quantitative assessments, demonstrating *i)* the high quality of the extracted multilingual alignments, *ii)* the novelty of the uncovered lexical semantic relations, and *iii)* the natural (rather than artificial) disambiguation power of the proposed multilingual approach.

2 Related Work

The problem of identifying the correct meaning of words depending on the context of occurrence represents one of the oldest tasks in the field of Natural Language Processing. The process of Word Sense Disambiguation hides a wide range of complexities, such that even after decades of technological advancement the current state of the art is still far from reaching more-than-good accuracy levels [26]. Many studies have already proved the advantages of a cross-lingual approach to Word Sense Disambiguation [8, 1, 10, 12]. The use of translations of a given word as sense labels avoid the need for manually created sense-tagged corpora and sense inventories. Moreover, a cross-lingual approach deals with the sense granularity problem: finer sense distinctions became truly relevant as far as they get lexicalized into different translations of the word [27]. However, existing works usually exploit either parallel texts or multilingual Wordnets, therefore relying on a intrinsically limited number of de-facto already built alignments.

Standard ways to encode lexical meaning are often based on explicit links between *words* and their possible *senses*, whereas words/senses are connected via paradigmatic relations (e.g., hypernymy, synonomy, antonymy, etc.), as in WordNet [30] and BabelNet [34]. 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. ConceptNet [42] 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 [28, 11] represent a similar kind of resource, which is more focused on 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. An emerging and extremely impactful approach to lexical semantics has been adopted by corpus-based and data-driven studies and technologies, which led to the creation of numeric (vectorial) encoding of lexical knowledge. This method is all centered on Harris' distributional assumption [17], i.e. words that occur in the same contexts tend to have similar

[3] https://www.media.mit.edu/

meanings. Well-known models include word embeddings [29, 36, 4], sense embeddings [19, 20, 25], and contextualized embeddings [39]. However, the relations holding between vector representations are not typed, nor are they organized systematically.

3 *k*-Multilingual Concepts

In this paper, we first propose the idea of *k-Multilingual Concept* (hereinafter MC^k), which consists of a concatenation of k lexical items referring to a single concept in k different languages. A MC^k can be described as a *pseudoword*, in line with the proposals put forward by [15] and [40], i.e., artificially-created words that can be used for different purposes (e.g., for the evaluation of Word Sense Induction systems [38]). In this instance, MC^ks are pseudowords that result from (and consist of) the alignment of multilingual, semantically equivalent lexical forms of a given concept. For example, if we consider the concept "*cat*" (as "*domestic cat*"), its $MC^{EN,IT}$ for the two languages English and Italian would be:

$$cat^{EN} \oplus gatto^{IT}$$

where the symbol $\oplus$ represents a simple concatenation operator. Similarly, we may extend the string by including other languages, adding e.g. a German equivalent word form. We would therefore obtain the following $MC^{EN,IT,DE}$:

$$cat^{EN} \oplus gatto^{IT} \oplus Katze^{DE}$$

A single MC^k is thus composed of k lexical forms, each one being linked to a specific language. However, the idea of a MC^k also presupposes that each of the k languages may have from zero to multiple lexicalizations of a given concept. The latter case would involve a synonymical set of words, whereas the former denotes what is referred to as lexical gap, i.e., concepts that lexicalize in one language but not in another. For example, the German reflexive verb *fremdschämen* in both Italian and English needs to be expressed with a periphrasis such as "*to feel embarrassed for someone*", since there is no lexical item with an equivalent meaning in the lexicons of either languages.

3.1 Lexical Gaps

Lexicalization is one of the linguistic devices available in natural languages for the integration of an item into the lexicon. This phenomenon typically involves a previously morphologically complex word that starts to acquire semantic and functional autonomy and behave as a single and independent lexical unit [43]. Being both a semantic notion and a process, it is gradient rather than categorical. Therefore, there can be different degrees of lexicalization. For example, the concept $\{leisure^{EN}, Freizeit^{DE}\}$ must be expressed in Italian through the multi-word expression *tempo libero*^{IT}. Despite being formed by two words, this expression nevertheless displays the same morphosyntactic and functional properties of the corresponding lexical forms in English and German. Thus, while *fremdschämen* is fully unlexicalized in Italian and English and generates a lexical gap, many lexical units such as *tempo libero*^{IT} or, e.g., English phrasal verbs represent lexical entries[4] albeit being slightly less-lexicalized than single-word units. Whenever the inventory of lexemes of a language does not include the full lexicalization of a given concept, such a lexical gap may create an empty value within a MC^k. This would be the case of *fremdschämen* or, e.g., of the Italian word *abbiocco* – which specifically denotes a feel of sleepiness caused by the digestion of an heavy meal. Thus, we will have:

$$\{\}^{EN} \oplus abbiocco^{IT} \oplus \{\}^{DE}$$

as $MC^{EN,IT,DE}$ associated with this concept. The idea of "*move body upright from sitting or lying*", instead, will be regularly encoded into the following $MC^{EN,IT,DE}$:

$$stand\ up^{EN} \oplus alzarsi^{IT} \oplus aufstehen^{DE}$$

3.2 Synonymous Words

A language may encode identical or similar semantic content into multiple word forms, causing instances of synonymy[5]. This will lead to a plurality of coordinated terms within the MC^k for a single concept. For example, if we only consider the English synonymical word forms *bike* and *bicycle*, we would have:

$$\{bike, bicycle\}^{EN} \oplus bicicletta^{IT} \oplus Fahrrad^{DE}$$

as $MC^{EN,IT,DE}$ associated with that single meaning[6].

3.3 Polysemous Words

Among the complex peculiarities of natural languages, that of polysemy (or *semantic ambiguity*) represents notoriously a challenging phenomenon for Natural Language Processing. Polysemy refers to the capacity for a word to convey multiple meanings, whereas the process of identification of its context-sensitive meaning is called disambiguation. However, each language features its own peculiar semantic system which, in turn, employs different formal encoding strategies. Therefore, by exploiting the different semantic (i.e. polysemous) behaviours of lexical items it is possible to disambiguate a given word by means of its semantic counterpart in another language.

[4] Therefore they are formally included in dictionaries, being considered as part of the lexicon by lexicographers.
[5] Yet synonymy, as a rule, is not complete equivalence - as we are reminded by [22].
[6] The same would apply for Italian and German synonyms for the concept *bicycle*.

The presented idea of MC^k is meant to represent a key instrument in this respect, since it is composed of a set of semantically equivalent lexical items that provide a quasi-monosemic (i.e. disambiguated) multilingual alignment. By providing a MC^k a context, or, more accurately, when a MC^k is generated through a defined linguistic context, their members will be indeed assigned a context-consistent meaning. Therefore, the MC^k will pinpoint a specific and unique concept. Finally, starting from the proven practice of leveraging multilingual word alignments to perform word disambiguation, we propose a novel methodology for automatically build them on a large scale without relying on already provided translations.

In the next section we will describe in detail the multilingual alignment mechanism that generates the MC^ks. This methodology, taken directly from [16], underpins the implementation of the MC^ks extraction as described thereafter.

4 Alignment Methodology

In this section, we present the alignment methodology used to automatically extract k-Multilingual Concepts from language-specific corpora.

4.1 Method and Languages Involved

As already performed in [16] we use three different languages in order to illustrate the building process of the multilingual resource. Thus, three European languages are involved in our work: English, German and Italian. The choice fell on these primarily because we are proficient in them, therefore we are able to properly handle and interpret the data. Furthermore, due to the very nature of the methodology, it was advisable to select a set of languages featuring a certain level of similarity in terms of shared lexical-semantic material. At the present stage, the alignment mechanism can be indeed effective and the results appreciable as long as the lexical-semantic systems of the languages involved reflect compatible cultural-linguistic backgrounds. A basic example will now help introduce the multilingual alignment mechanism. Consider the concept "*wool*" (as "*textile fiber obtained from sheep and other animals*") and the tree word forms $\{wool^{EN}, lana^{IT}, Wolle^{DE}\}$, constituting the following $MC^{EN,IT,DE}$:

$$wool^{EN} \oplus lana^{IT} \oplus Wolle^{DE}$$

The so conceived *head* concept represents our starting point from which a linguistic context will be generated. Hence, we may represent it also as:

$$MC^{EN,IT,DE}_{wool-textile\ fiber}$$

For each of the three word forms that compose the $MC^{EN,IT,DE}$ *head* we retrieve a set of semantically related words of different types (nouns, modifiers, verbs) in terms of paradigmatic (e.g. synonyms) and syntagmatic (e.g. co-occurrences) relations. We thus obtain three different lists of *head*-related words, one for each of the three languages. Table 1 provides a small excerpt of such unordered lists.

wool^{EN}	**lana**^{IT}	**Wolle**^{DE}
sheep	*cotone*	*Schal*
cotton	*Biella*	*spinnen*
synthetic	*sintetica*	*Baumwolle*
spin	*sciarpa*	*Rudolf*
scarf	*pecora*	*synthetisch*
mitten	*filare*	*Schafe*

Table 1. Unordered lists of single-language related words for $MC^{EN,IT,DE}_{wool-textile\ fiber}$.

The retrieved terms in the lists may be still ambiguous, since they are related to a word form rather than to a contextually defined concept. Thus, the lexical data in the lists are subsequently compared and filtered by means of a translation step, in order to select only the semantic items that occur in all the lists, i.e., those shared by the three languages. The resulting words are thus aligned with their semantic counterparts, as shown in Table 2.

wool^{EN}		**lana**^{IT}		**Wolle**^{DE}
sheep	$\oplus$	*pecora*	$\oplus$	*Schafe*
cotton	$\oplus$	*cotone*	$\oplus$	*Baumwolle*
synthetic	$\oplus$	*sintetica*	$\oplus$	*synthetisch*
spin	$\oplus$	*filare*	$\oplus$	*spinnen*
scarf	$\oplus$	*sciarpa*	$\oplus$	*Schal*

Table 2. Examples of aligned concept-related words for $MC^{EN,IT,DE}_{wool-textile\ fiber}$.

As can be noted, by combining, e.g., the lexical form *to spin* with the Italian word *filare* and the German *spinnen* - which, among others, encode one of the possible senses of *spin* - we would obtain the following $MC^{EN,IT,DE}$:

$$spin^{EN} \oplus filare^{IT} \oplus spinnen^{DE}$$

Once aligned, the three previously polysemous lexical forms constitute a $MC^{EN,IT,DE}$ that refers to a specific and unique conceptualization, i.e., "*turn fibers into thread*". The resulting list of $MC^{EN,IT,DE}$ for the head concept $MC^{EN,IT,DE}_{wool-textile\ fiber}$ provides an encoding of lexical knowledge linked to the seed concept which is i) *unbiased*, since the filtering step enables to avoid language-bounded material by including only items that are shared by all three languages; ii) *diversified*, since it consist of both paradigmatic and syntagmatic lexical relations for three different POS.

4.2 Automatic Extraction of MC^ks

We built a data ingestion process that automatically outputs MC^ks, using as mentioned above k=3 languages: English (EN), Italian (IT) and German (DE). To start an automatic MC^k extraction process for a generic concept C the first requirement is to have a seed, i.e., a MC^k *head* that is constituted by k word forms representing C, one for each language. Since a generic concept C may present language-related issues (e.g. lexical gaps - see Section 3.1), we retrieve MC^k *heads* directly from BabelNet synsets. In particular, given a BabelNet synset for a concept C, we select a maximum of 3 *high-quality* lexicalizations[7] for each language. If BabelNet does not provide at least one high quality lexicalization for each language, we rely on Open Multilingual Wordnet project [6] to look for English and Italian lexicalizations and OdeNet [41] for German ones, while Collaborative InterLingual Index (CILI) [7] serves as a link between the two to retrieve the shared synset. The obtained word forms in the three languages will constitute the MC^k *head* around which the procedure will autonomously extract the multilingual knowledge around C.

Once the MC^k *head* has been formed, we use Sketch Engine [24], a corpus management engine, to obtain lists of words related to each single word form that makes up the MC^k *head*, as shown in the example in Table 1. We employ three families of non-semantically annotated large corpora to search for related words in the three languages: the TenTen corpora containing 10+ billion words of generic web content [23], the TJSI corpora composed of news articles [44][8] and the EUR-Lex legal corpora [2]. Then, we merge the retrieved related words in the three target languages obtaining three lists (hereinafter *EN*-list, *IT*-list and *DE*-list), each divided into four categories: *i)* similar nouns, *ii)* co-occurring nouns, *iii)* co-occurring adjectives and *iv)* co-occurring verbs. Finally, we assign a weight to each related word by directly importing the built-in scores of Sketch Engine tools, that are based on the Dice coefficient, as detailed in [24].

To obtain the the MC^ks alignments like those shown in Table 2 we search for cross-match translations using the PanLex API[9], which is focused on words rather than on sentences, and the Google Translate API[10]. Specifically, we take each related word, category by category, from the *EN*-list and query the API to get their possible translations into Italian, ordered by confidence. If we find a match between such translations and a related word in the *IT*-list of equal category, we form a pair $<rw^{EN}, rw^{IT}>$. Once all possible pairs have been identified, we repeat the procedure starting from all rw^{EN}s to find matches within the *DE*-list of the same category, thus obtaining triplets $<rw^{EN}, rw^{IT}, rw^{DE}>$. A final verification is performed by testing the correct correspondence between each $<rw^{IT}, rw^{DE}>$ pair, through the same cross-match translation process. If this

[7] BabelNet high-quality lexicalizations are those word forms that are not marked as resulting from an automatic translation.

[8] TJSI versions used: English (60+ billion words), Italian (8.4+ billion words), German (6.9+ billion words).

[9] https://dev.panlex.org/api/.

[10] https://cloud.google.com/translate.

step fails, the whole triplet will be marked as *weak*. Otherwise, the successful alignment will be considered as *strong* and will constitute a $MC^{EN,IT,DE}$. We finally assign a score to each $MC^{EN,IT,DE}$ by averaging the SketchEngine scores of the three related words.

As last step, we associate BabelNet synsets (always those directly linked to WordNet synsets, if present) and WordNet synsets to the alignments. Specifically, we find the n synsets that have all the given three word forms in the three languages. One of the following three cases may hence occur: *i)* $n = 1$, meaning that the $MC^{EN,IT,DE}$ corresponds to a completely disambiguated concept; *ii)* $n > 1$, when multiple synsets may be associated with a single $<rw^{EN}$, rw^{IT}, $rw^{DE}>$ triplet; *iii)* $n = 0$, in case no existing BabelNet synset or WordNet synset actually connects the three word forms. It is interesting to note that the last two cases cover different situations, such as a missing synset econding a specific concept ($n = 0$, e.g. significant for sense induction) or overlapping synsets ($n > 1$, e.g. useful for sense clustering).

5 The *MultiAligNet* Resource

The k-Multilingual Concept model and the automatic extraction method we developed allowed us to create an original lexical-semantic resource, which we refer to as *MultiAligNet*. To date, the resource is publicly available[11] and contains the extracted knowledge referring to 1047 synsets that we used as *heads*, which corresponds to a total of 21514 automatically-built MC^ks over the three languages. Future updates will be made available within the same repository. The selection of *head* concepts has been performed carefully. First, we manually selected 100 concepts by inspecting basic vocabularies of each of the three languages[12], covering different semantic categories and characteristics such as the degrees of polysemy and abstractness. Then we automatically retrieved the 750 most frequent and 200 rare concepts in SemCor [31], one of the most used sense-annotated corpora to train supervised WSD systems. Finally, we randomly-picked a set of polysemous words referring to more than 50 synsets in total. The *MultiAligNet* resource is available in two different formats, as described below.

5.1 Distributional Representation

Our resource can be displayed through a vectorial representation of the k-Multilingual Concepts. In particular, synsets are represented as vectors whose dimensions point to the synsets linked to the alignments (see Section 4.2 for details). Such distributional version of the resource is different from standard word- and sense-embedding technologies, since features are conceptual (being

[11] https://github.com/vloverar/multialignet

[12] For EN: iWebCorpus, The Oxford Dictionary https://www.english-corpora.org/iweb, https://www.oxfordlearnersdictionaries.com/wordlists/oxford3000-5000; for IT: *NvdB* https://www.dropbox.com/s/mkcyo53m15ktbnp/nuovovocabolariodibase.pdf; for DE: [45].

connected to real synsets). This is similar to what happens with Explicit Semantic Analysis (ESA) [14], Salient Semantic Analysis (SSA) [18] and others [9]. This version may be employed in semantic similarity tasks and, generally, in the context of Explainable AI research.

5.2 Knowledge Graph

Similarly to other lexical-semantic resources, our model reflects a deep interconnection of term- and concept-based items, which makes it well-suited for a graph-based knowledge encoding. We provide a knowledge graph relying on the Neo4j[13] database open technologies and libraries. In the graph model we

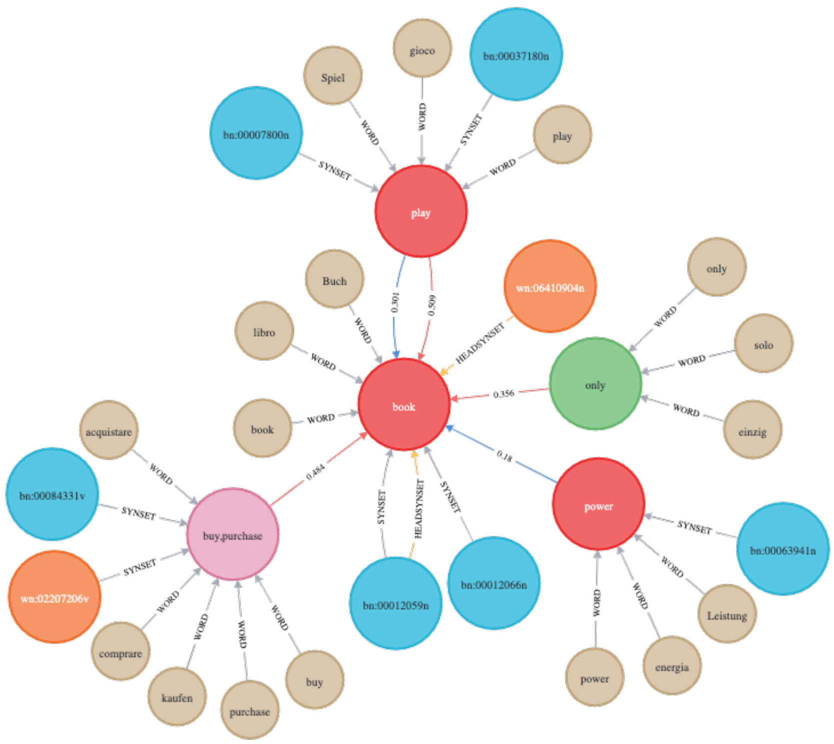

Fig. 1. Illustrative excerpt of *MultiAligNet* graph around the $MC^{EN,IT,DE}_{book-written\ work}$ *head*. Red, pink and green circles represent *align*-nodes for nouns, verbs and adjectives respectively (for space requirements, only the English word forms are displayed). Beige, blue and orange ones represent *word*-, *babel synset*- and *wordnet synset*-nodes.

employ four types of nodes, namely *i) word*-nodes, *ii) babel synset*-nodes, *iii)*

[13] https://neo4j.com

wordnet synset-nodes and *iv) align*-nodes (further typed with POS tags). While the first three enable standard access features for words- and synsets-centered queries (as in WordNet and BabelNet), *align*-nodes represent a novel type of information, specifically hinged on the MC^k multilingual concatenations of terms. The released *MultiAligNet* knowledge graph contains 72,469 nodes, interconnected by 387,273 relations. Figure 1 shows an excerpt of the graph around the $MC^{EN,IT,DE}_{book-written\ work}$ *head.*

6 Extraction Results and Evaluation

Starting from our selected concepts (1,047 *heads*), we automatically extracted 21,514 multilingual alignments (MC^ks). Among them, 9,007 (41.86%) do not present any available linking to either WordNet or BabelNet synsets (for the latter, considering only the *high quality* lexicalizations) whereas 1,045 have an available linking only to *low-quality* lexicalization in BabelNet. Finally, 7,962 triplets (37.01%) present no available linking to either WordNet or BabelNet, considering both *high-* and *low quality* lexicalizations. This latter data refers to totally novel lexical knowledge compared to the two reference resources.

In this section, we first report the results of a qualitative assessment of such generated knowledge. We then outline a quantitative evaluation reflecting the impact of MC^ks in uncovering novel semantic relations with respect to a state-of-the-art existing repository (i.e. BabelNet) without making use of any Word Sense Disambiguation (WSD) system.

6.1 MC^ks Novelty and Quality Assessment

7,962 MC^ks out of 21,514 present no available linking to either WordNet or Babelnet synsets. This means that the system managed to retrieve novel lexical knowledge quantifiable as 7,962 alignments related to 1,047 head concepts. We then manually evaluated the quality of these new MC^ks in order to assess whether they consist of actually valid three-lingual lexicalizations of single concepts. In particular, we manually checked a randomized subset of 250 triplets. The manual check was performed by assessing the semantic equivalence of each MC^k, thus validating the translations of each word of the alignment into the other two by using bilingual dictionaries[14]. We assessed both translation directions for each word pair ($<rw^{EN}, rw^{IT}>$; $<rw^{EN}, rw^{DE}>$; $<rw^{DE}, rw^{IT}>$). The semantic equivalence assessment task showed that a total of 235 out of 250 MC^k (93.6%) were indeed accurate. Finally, we measured the amount of novel connections retrieved by MultiAligNet with respect to the BabelNet knowledge graph. Interestingly, 264,813 links between alignments (out of 290,730) are not present in BabelNet.

[14] The annotator who performed the evaluation is however a native Italian speaker with a minimum of C1 both English and German proficiency level. Therefore, the evaluation is assured by a solid accuracy.

6.2 MC^ks Disambiguation Power

The MC^k model enables a peculiar encoding of lexical knowledge which lies between the high polysemy of words and the static nature of predefined word senses. Therefore, we aim to concretely measure to what extent MC^ks can reduce single-language word ambiguity without relying on any WSD method. Hence, for each polysemous word w^L in a given language L, we can count its possible senses $ns(w^L) \geq 2$, as well as the resulting senses linked to the k-multilingual concept $ns(MC^k_{w^L})$. Note that $ns(w^L)$ is always greater than or equal to $ns(MC^k_{w^L})$. We can compute a disambiguation power (dp) index for a single word w^L as follows:

$$dp(w^L, MC^k_{w^L}) = \frac{ns(w^L) - max(1, ns(MC^k_{w^L}))}{ns(w^L) - 1}$$

Note that since MC^ks may not be linked to any synset (as mentioned in Section 4.2), the *max* function forces to 1 the value of the subtrahend. The range of the dp is $[0, 1]$ where 0 means no disambiguation and 1 maximum disambiguation (this latter case occurs whenever all senses $ns(w^L)$ got reduced to a single MC^k sense (i.e. $ns(MC^k_{w^L}) = 1$)). In order to obtain an overall MC^k dp-index for a set of target words in a language L, we can compute an average score as follows:

$$dp^L = \frac{1}{|w^L|} \sum_{\forall w^L} dp(w^L, MC^k_{w^L})$$

Table 3 shows the dp index for the three languages. Impressively, MC^ks considerably reduced single-language word ambiguity in all three languages. In particular, for the *EN*- and *IT*-ambiguous lexical entries, the proposed alignment was able to reduce their polysemy by 85%. This demonstrates the high potential of the MC^k model in encoding mostly-unambiguous lexical knowledge without relying on fixed sense repositories.

Language	n. of ambiguous words	*dp*-index
EN	9480	0.851
IT	7395	0.852
DE	4866	0.756

Table 3. Disambiguation power (dp) index for the three languages *EN*, *IT*, *DE*.

7 Conclusion and Future Work

In this paper, we proposed a novel encoding method for the representation of lexical-semantic knowledge based on the idea of k-Multilingual Concept (MC^k). The developed methodology allows the automatic alignment of semantically equivalent words in k different languages as occurring in a determined linguistic context. The resulting alignments result in a cross-lingual encoding of unbiased and multifaceted lexical knowledge, in terms of empirically determined conceptual links consisting of syntagmatic and paradigmatic lexical relations.

We then released *MultiAligNet*, an original resource containing, to date, more than 21k automatically-extracted MC^ks on a heterogeneous selection of concepts in English, Italian and German. We thus evaluated the resource by means of both qualitative and quantitative assessments on the data retrieved. Results demonstrate the validity of the method concerning its ability to retrieve (i) unbiased lexical knowledge (ii) diversified lexical relations (iii) novel lexical material as compared to existing resources (BabelNet and WordNet). Finally, the proposed model enabled a natural (multilingual) disambiguation mechanism for words without the help of sense repositories or parallel texts. In future work, we aim to continuously extend the resource by covering more concepts and languages, fostering novel research on different tasks such as enrichment, disambiguation and induction of senses in existing repositories.

References

1. Apidianaki, M.: LIMSI : Cross-lingual word sense disambiguation using translation sense clustering. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 178–182. Association for Computational Linguistics, Atlanta, Georgia, USA (Jun 2013), https://aclanthology.org/S13-2032
2. Baisa, V., Michelfeit, J., Medveď, M., Jakubíček, M.: European union language resources in sketch engine. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 2799–2803 (2016)
3. Barba, E., Procopio, L., Navigli, R.: Consec: Word sense disambiguation as continuous sense comprehension. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 1492–1503 (2021)
4. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
5. Bond, F., Foster, R.: Linking and extending an open multilingual wordnet. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1352–1362 (2013)
6. Bond, F., Foster, R.: Linking and extending an open multilingual Wordnet. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1352–1362. Association for Computational Linguistics, Sofia, Bulgaria (Aug 2013), https://aclanthology.org/P13-1133
7. Bond, F., Vossen, P., McCrae, J., Fellbaum, C.: CILI: the collaborative interlingual index. In: Proceedings of the 8th Global WordNet Conference (GWC). pp. 50–57. Global Wordnet Association, Bucharest, Romania (27–30 Jan 2016)
8. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L.: Word-sense disambiguation using statistical methods. In: 29th Annual Meeting of the Association for Computational Linguistics. pp. 264–270. Association for Computational Linguistics, Berkeley, California, USA (Jun 1991)
9. Camacho-Collados, J., Pilehvar, M.T., Navigli, R.: Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. Artificial Intelligence **240**, 36–64 (2016)

10. Chan, Y.S., Ng, H.T.: Scaling up word sense disambiguation via parallel texts. In: Proceedings of the 20th National Conference on Artificial Intelligence - Volume 3. p. 1037–1042. AAAI'05, AAAI Press (2005)
11. Devereux, B.J., Tyler, L.K., Geertzen, J., Randall, B.: The cslb concept property norms. Behavior research methods **46**(4), 1119–1127 (2014)
12. Diab, M.T., Resnik, P.: Word Sense Disambiguation within a Multilingual Framework. Ph.D. thesis, USA (2003), aAI3115805
13. Edmonds, P., Kilgarriff, A.: Introduction to the special issue on evaluating word sense disambiguation systems. Natural Language Engineering **8**(4), 279–291 (2002)
14. Gabrilovich, E., Markovitch, S., et al.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: IJcAI. vol. 7, pp. 1606–1611 (2007)
15. Gale, W.A., Church, K.W., Yarowsky, D.: Work on statistical methods for word sense disambiguation. In: Working Notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language. vol. 54, p. 60 (1992)
16. Grasso, F., Di Caro, L.: A methodology for large-scale, disambiguated and unbiased lexical knowledge acquisition based on multilingual word alignment. In: Fersini, E., Passarotti, M., Patti, V. (eds.) Proceedings of the Eighth Italian Conference on Computational Linguistics, CLiC-it 2021, Milan, Italy, January 26-28, 2022. CEUR Workshop Proceedings, vol. 3033. CEUR-WS.org (2021)
17. Harris, Z.S.: Distributional structure. Word **10**(2-3), 146–162 (1954)
18. Hassan, S.H., Mihalcea, R.: Semantic relatedness using salient semantic analysis. In: Twenty-Fifth AAAI Conference on Artificial Intelligence (2011)
19. Huang, E.H., Socher, R., Manning, C.D., Ng, A.Y.: Improving word representations via global context and multiple word prototypes. In: Proc. of ACL. pp. 873–882 (2012)
20. Iacobacci, I., Pilehvar, M.T., Navigli, R.: SensEmbed: learning sense embeddings for word and relational similarity. In: Proceedings of ACL. pp. 95–105 (2015)
21. Ion, R., Tufis, D.: Multilingual word sense disambiguation using aligned wordnets. Romanian Journal of Information Science and Technology Volume **7**, 183–200 (01 2004)
22. Jakobson, R.: 14. On Linguistic Aspects of Translation, pp. 144–151. University of Chicago Press (2012)
23. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlỳ, P., Suchomel, V.: The tenten corpus family. In: 7th International Corpus Linguistics Conference CL. pp. 125–127 (2013)
24. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The sketch engine: Ten years on. The Lexicography **1(1)**, 7–36 (2014)
25. Kumar, S., Jat, S., Saxena, K., Talukdar, P.: Zero-shot word sense disambiguation using sense definition embeddings. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5670–5681 (2019)
26. Lacerra, C., Bevilacqua, M., Pasini, T., Navigli, R.: Csi: A coarse sense inventory for 85% word sense disambiguation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 8123–8130 (2020)
27. Lefever, E., Hoste, V.: SemEval-2013 task 10: Cross-lingual word sense disambiguation. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 158–166. Association for Computational Linguistics, Atlanta, Georgia, USA (Jun 2013), https://aclanthology.org/S13-2029

28. McRae, K., Cree, G.S., Seidenberg, M.S., McNorgan, C.: Semantic feature production norms for a large set of living and nonliving things. Behav. r. m. **37**(4), 547–559 (2005)
29. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
30. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
31. Miller, G.A., Chodorow, M., Landes, S., Leacock, C., Thomas, R.G.: Using a semantic concordance for sense identification. In: Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994 (1994)
32. Morris, J., Hirst, G.: Non-classical lexical semantic relations. In: Proceedings of the Computational Lexical Semantics Workshop at HLT-NAACL 2004. pp. 46–51. Association for Computational Linguistics, Boston, Massachusetts, USA (May 2 - May 7 2004), https://aclanthology.org/W04-2607
33. Navigli, R.: Word sense disambiguation: A survey. ACM computing surveys (CSUR) **41**(2), 1–69 (2009)
34. Navigli, R., Ponzetto, S.P.: BabelNet: Building a very large multilingual semantic network. In: Proc. of ACL. pp. 216–225. Association for Computational Linguistics (2010)
35. Palmer, M., Dang, H.T., Fellbaum, C.: Making fine-grained and coarse-grained sense distinctions, both manually and automatically. Nat.Lan.Eng. **13**(02), 137–163 (2007)
36. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP. vol. 14, pp. 1532–43 (2014)
37. Petricca, P.: SEMANTICA. Forme, modelli, problemi (10 2019)
38. Pilehvar, M.T., Navigli, R.: A large-scale pseudoword-based evaluation framework for state-of-the-art word sense disambiguation. Computational Linguistics **40**(4), 837–881 (2014)
39. Scarlini, B., Pasini, T., Navigli, R.: 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 (2020)
40. Schütze, H.: Dimensions of meaning. In: SC. pp. 787–796 (1992)
41. Siegel, M., Bond, F.: OdeNet: Compiling a GermanWordNet from other resources. In: Proceedings of the 11th Global Wordnet Conference. pp. 192–198. Global Wordnet Association, University of South Africa (UNISA) (Jan 2021), https://aclanthology.org/2021.gwc-1.22
42. Speer, R., Chin, J., Havasi, C.: Conceptnet 5.5: An open multilingual graph of general knowledge (2017)
43. Thomas, C.: Lexicalization in Generative Morphology and Conceptual Structure, pp. 45–65. Edinburgh University Press (2013)
44. Trampuš, M., Novak, B.: Internals of an aggregated web news feed. In: Proceedings of 15th Multiconference on Information Society. pp. 221–224 (2012)
45. Tschirner, E.: Deutsch nach Themen: Grund-und Aufbauwortschatz: Deutsch als Fremdsprache nach Themen-Lernwörterbuch. Cornelsen, Berlin (2016)