Conception: Multilingually-Enhanced, Human-Readable Concept Vector Representations

Simone Conia and **Roberto Navigli**
Sapienza NLP Group
Department of Computer Science
Sapienza University of Rome
{conia,navigli}@di.uniroma1.it

Abstract

To date, the most successful word, word sense, and concept modelling techniques have used large corpora and knowledge resources to produce dense vector representations that capture semantic similarities in a relatively low-dimensional space. Most current approaches, however, suffer from a monolingual bias, with their strength depending on the amount of data available across languages. In this paper we address this issue and propose Conception, a novel technique for building language-independent vector representations of concepts which places multilinguality at its core while retaining explicit relationships between concepts. Our approach results in high-coverage representations that outperform the state of the art in multilingual and cross-lingual Semantic Word Similarity and Word Sense Disambiguation, proving particularly robust on low-resource languages. Conception – its software and the complete set of representations – is available at https://github.com/SapienzaNLP/conception.

1 Introduction

Word vector representations, in particular dense representations or word embeddings (Mikolov et al., 2013a; Pennington et al., 2014; Bojanowski et al., 2017), play a key role in a wide range of tasks, including Text Similarity (Kenter and de Rijke, 2015; Nguyen et al., 2019), Word Sense Disambiguation (Iacobacci et al., 2016; Raganato et al., 2017a), Semantic Role Labeling (He et al., 2017; Marcheggiani et al., 2017; Conia et al., 2020), Question Answering (Zhou et al., 2015) and Machine Translation (Mikolov et al., 2013b; Bahdanau et al., 2015). This is especially the case when they are used as the underlying input representation. Word embedding techniques map each word to a relatively low n-dimensional space where two semantically or syntactically similar words lie close together. Due to their latent nature, however, most embeddings are commonly considered to be uninterpretable (Levy and Goldberg, 2014) as the properties captured by each dimension are often unclear. More recent studies have shed some light on their interpretability (Rothe and Schütze, 2016; Senel et al., 2018; Wallace et al., 2019) or included interpretability directly in the learning process (Park et al., 2017; Koç et al., 2018), but the opaqueness of dense vectors is still a key reason why research has not completely given up on sparse representations (Faruqui et al., 2015; Derby et al., 2018).

Moreover, most current embedding techniques rely on large corpora which are often available in few languages, such as English or Chinese, strongly limiting their robustness on low-resource languages (Speer and Lowry-Duda, 2017). In an attempt to solve this issue, researchers turned to multilingual word representations by making use of parallel vocabularies (Mikolov et al., 2013b; Ammar et al., 2016; Smith et al., 2017), exploiting multilingual knowledge graphs (Speer et al., 2017), and exploring unsupervised methods to align monolingual embeddings in a single shared distributional space (Conneau et al., 2017) or to directly learn multilingual embeddings (Chen and Cardie, 2018).

Nevertheless, a well-known pitfall of both monolingual and multilingual word representations is the so-called *meaning conflation deficiency problem* (Camacho-Collados and Pilehvar, 2018): a word may be ambiguous, that is, it may have multiple meanings, but those possibly unrelated meanings cannot

Proceedings of the 28th International Conference on Computational Linguistics, pages 3268–3284
Barcelona, Spain (Online), December 8-13, 2020

be told apart since they are conflated into a single representation. As a result, contextualized word representations have garnered attention (Melamud et al., 2016), enjoying great success in the form of pretrained language models like BERT (Devlin et al., 2019) or XLM (Conneau and Lample, 2019). At the same time, modelling techniques for individual word senses, concepts and named entities have also gained traction (Camacho-Collados et al., 2016; Scarlini et al., 2020a), though their integration into downstream NLP applications is still subject of ongoing investigations (Li and Jurafsky, 2015; Pilehvar et al., 2017).

The requirement of massive amounts of training data and the lack of interpretability hinder most of the above-mentioned approaches. To address these limits, we introduce Conception, a novel knowledge-based technique for modelling concepts and named entities through concepts and named entities. Our approach places multilinguality at its core by leveraging the mutually-reinforcing information coming from different languages, enabling seamless and robust cross-lingual scaling, while also providing explicit and easily interpretable semantic dimensions. In contrast to most word-based embeddings, in Conception:

i. each component in a vector represents a (weighted) concept or named entity, therefore making our representations fully interpretable;

ii. vector representations are explicitly linked to BabelNet (Navigli and Ponzetto, 2012a), a multilingual semantic network which provides coverage for words and multiword expressions in 284 languages;

iii. each concept and named entity is defined as a language-independent unit, so the same representation can be used across languages.

We evaluate Conception on multilingual and cross-lingual Semantic Word Similarity, finding that our approach outperforms supervised, unsupervised and knowledge-based state-of-the-art techniques for both sparse and dense vector representations. Furthermore, we show that these improvements translate into the downstream task of Word Sense Disambiguation, where Conception surpasses the state of the art among supervised and knowledge-based techniques, showing that our semantics-first representations contain meaningful information even when compared against BERT-based techniques.

2 Related Work

Multilingual word embeddings. The advantages of multilinguality in representation learning were first noticed by Mikolov et al. (2013b), who exploited similarities in the structures of the distributional spaces of different languages to learn cross-lingual word embeddings by taking advantage of purposely-built parallel vocabularies. Since then, multilinguality has become increasingly important in learning robust representations: Faruqui and Dyer (2014) used canonical correlation analysis to project independently-constructed distributional spaces for two languages onto a common space; Ammar et al. (2016) extended previous work to over fifty languages; Smith et al. (2017) reduced the need for bilingual supervision by compiling a pseudo-dictionary from the identical strings that appear in two languages; Jawanpuria et al. (2019) proposed a geometric approach to embedding alignment that leverages language-specific transformations; Singhal et al. (2019) learned multilingual word embeddings from image-text data. While these methods still require annotated cross-lingual data or parallel vocabularies, Conneau et al. (2017) and Artetxe et al. (2018) found success by employing unsupervised methods and adversarial training.

Contextualized word embeddings. The above-mentioned approaches produce word-level representations that are independent of the specific context a word appears in, and such "static" representations often show a strong bias towards the most frequent sense of a word. Instead, context-aware word representation techniques, such as context2vec (Melamud et al., 2016) or ELMo (Peters et al., 2018), dynamically create a representation for a word in a sentential or documental context. Contextualized embeddings witnessed a dramatic rise in popularity thanks to the advent and wide availability of language models pretrained on massive amounts of text, such as BERT (Devlin et al., 2019), immediately followed by multilingual language models, such as m-BERT and XLM (Conneau and Lample, 2019). Contextualized word embeddings are able to capture the many facets of a polysemous word in a context (Pilehvar and Camacho-Collados, 2019), but their implicitly encoded meanings are still disconnected from human-curated knowledge bases, even if more recent efforts showed promising results in imparting structured semantic knowledge to contextualized representations (Peters et al., 2019; Levine et al., 2019).

Knowledge-enhanced representations. Alternative approaches implement multilinguality by making use of multilingual encyclopedic resources, such as Wikipedia (Al-Rfou et al., 2013), or multilingual knowledge graphs such as Open Multilingual WordNet (Bond and Foster, 2013), ConceptNet (Speer et al., 2017), or BabelNet (Navigli and Ponzetto, 2012a). A prominent example of knowledge-enhanced word embeddings is Conceptnet Numberbatch (Speer et al., 2017), which retrofits word2vec (Mikolov et al., 2013a) and GloVe (Pennington et al., 2014) word representations to the ConceptNet graph, achieving state-of-the-art results in Semantic Word Similarity (Speer and Lowry-Duda, 2017).

A further step towards semantic representations involves modelling individual word senses as vectors which are explicitly linked to a knowledge resource. Early approaches to sense embeddings adapted existing work to project words and word senses onto a shared distributional space (Iacobacci et al., 2015; Iacobacci and Navigli, 2019), while more recent studies exploited the inner states of pretrained language models (Loureiro and Jorge, 2019; Scarlini et al., 2020a). Instead of modelling language-specific units like words or senses, NASARI (Camacho-Collados et al., 2016) represents language-independent concepts using sparse lexical vectors. Its most notable shortcoming, however, is that each lexical vector is built from a single source language, and therefore each concept has a separate representation depending on the source language of choice. While the NASARI lexical vectors provide language-specific representations for language-independent concepts, Camacho-Collados et al. (2016) also proposed a "unified" variant where the vector dimensions are concepts obtained by semantically clustering the words of the corresponding lexical vector based on the hypernymy relation. However, such representations start from a single language, and therefore do not exploit the multilingual content available in resources such as BabelNet, and are not interrelated to each other. With Conception, we tackle all these issues and propose an integrated, multilingually-enhanced representation of concepts and entities.

3 Preliminaries

Conception relies on the concept inventory of BabelNet and the lexical vectors of NASARI to build its representations, which we introduce hereafter.[1]

BabelNet (Navigli and Ponzetto, 2012a) is a multilingual semantic network that brings together heterogeneous resources, such as Wikipedia, WordNet, and Open Multilingual WordNet, with 284 languages supported in the current version 4.0. Each node in the BabelNet graph represents a concept or named entity and is defined as a multilingual synset, i.e., the set of synonymous lexicalizations used in different languages to express the same concept or named entity.[2] For example, the concept MOTOR VEHICLE is defined as the multilingual synset containing the terms { car_{EN}, $\text{motorcar}_{\text{EN}}$, coche_{ES}, $\text{voiture}_{\text{FR}}$, $\text{macchina}_{\text{IT}}$, ..., 自動車$_{\text{JP}}$}. In BabelNet, synsets are connected to other synsets through a variety of relations, from hypernymy (generalization or is-a) to hyponymy (specialization or has-kind), from meronymy (part-whole) to antonymy (opposite-of) and general relatedness relations extracted from Wikipedia page links, among others. While some different relation types may arguably be considered more important than others, for the sake of simplicity, we do not distinguish between synset relation types.

NASARI (Camacho-Collados et al., 2016), as previously mentioned, represents a concept c in the form of a sparse vector $\mathbf{v}_c^l$ whose components are the weights of lexical items (words and multiword expressions) expressed in a given language l. The weight of a lexical item w is computed as its lexical specificity (Lafon, 1980) in the subcorpus of Wikipedia articles which define c and its related concepts

[1] We note that our approach does not rely on any BabelNet- or NASARI-specific feature, i.e., in principle, BabelNet can be seamlessly replaced with another multilingual semantic network such as Open Multilingual WordNet.

[2] From now on, we refer to concepts, named entities and synsets interchangeably.

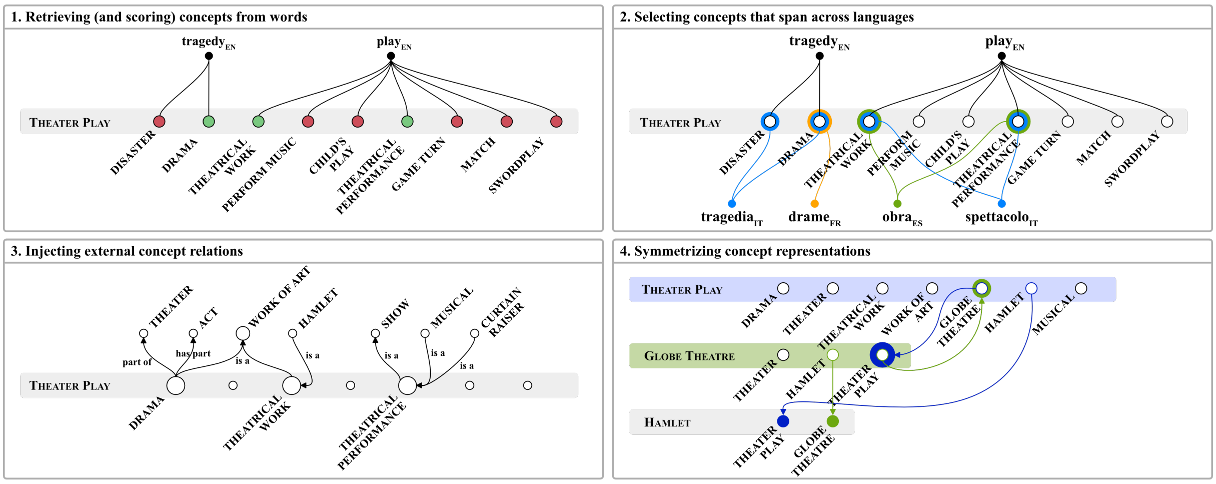

Figure 1: A (simplified) visualization of the Conception algorithm. Top left: retrieving all the possible meanings of each lexical item in the NASARI lexical vector (Section 4.1). Top right: selecting the most relevant concepts through cross-lingual disambiguation (Section 4.2). Bottom left: exploiting the concept relations in BabelNet to inject external knowledge in the representations (Section 4.3). Bottom right: enforcing symmetric relations, i.e., if a concept c_a appears in the representation of c_b, then c_b should also appear in the representation of c_a (Section 4.4).

using language l. Lexical specificity is based on the hypergeometric distribution over word frequencies in such corpora and is computed as $\mathbf{v}_c^l[w] = \text{spec}(w) = -\log P(X \geq f; T, t, F, f)$, where T and t are the sizes of Wikipedia and the subcorpus, respectively, and F and f are the frequencies of w in the two respective corpora. For each language l in Wikipedia, NASARI can produce a distinct representation $\mathbf{v}_c^l$ of a concept c. However, since vocabularies of different languages are mostly non-overlapping, the components or lexical items of $\mathbf{v}_c^{l'}$ and $\mathbf{v}_c^{l''}$ cannot be directly compared across any two languages l' and l''. Notably, $\mathbf{v}_c^{l'}$ may include knowledge that is missing from $\mathbf{v}_c^{l''}$ and, at the same time, the lexical items of $\mathbf{v}_c^{l'}$ can help disambiguate the lexical items of $\mathbf{v}_c^{l''}$, as observed by Navigli and Ponzetto (2012b).

4 Conception

The key innovation we put forward is that, with Conception, multilinguality is an integral part of the learning process: instead of deriving concept representations from within a single language, we leverage the mutually-reinforcing information available across languages to create human-readable and language-independent concept-level representations. Our approach results in representations where each concept c is described by a vector where each dimension corresponds to a concept: a larger magnitude for the i-th component denotes a stronger relation between c and the i-th concept c_i, that is, each concept is described by the concepts it is most related to.

By modelling individual concepts rather than words, Conception does not suffer from the conflation of senses that affects word-level representations, such as word2vec and GloVe. At the same time, since concepts are language-independent units, using them as the dimensions of our representations addresses the language-specificity issue of other sparse representations such as the NASARI lexical vectors (see Section 3).

In the remainder of this Section, we describe the four steps of Conception (a running example, discussed in what follows, is shown in Figure 1).

4.1 Retrieving concepts from any language

Starting from the NASARI lexical vectors (available in 5 languages, namely English, French, German, Italian, and Spanish), the first step of Conception obtains sparse vectors whose dimensions are concepts instead of words or multiword expressions. Given a concept c and a language l, let $\mathbf{v}_c^l$ be the lexical vector representing c in l. For each such vector $\mathbf{v}_c^l$ and for each non-null lexical item w in $\mathbf{v}_c^l$, we consider each concept c' that has a lexicalization w in l according to BabelNet, and score the relevance of c' with respect to c as:

$$\textsc{Score}(c'|\mathbf{v}_c^l, w) = \textsc{W-Rank}(\mathbf{v}_c^l, w)^{-1}$$

where $\textsc{W-Rank}(\mathbf{v}_c^l, w)$ is the ranking of the lexical item w among the components of $\mathbf{v}_c^l$ sorted by decreasing magnitude. The language-independent aggregated score of c' with respect to c is computed across the set of source languages L as follows:

$$\textsc{Score}(c'|c) = \sum_{l \in L} \sum_{w \in c'} \textsc{Score}(c'|\mathbf{v}_c^l, w)$$

As a result, for a given concept c, we can create an initial semantic vector $\mathbf{v}_c$ whose components are $\mathbf{v}_c[c'] = \textsc{Score}(c'|c)$, for each c' in BabelNet. This step does not disambiguate lexical items, so a score is assigned to all the concepts collected from the items of any lexical vector $\mathbf{v}_c^l$, leading to noisy representations that include undesired concept relations. For example, if the English lexical representation for the THEATER PLAY concept includes the lexical item *play*$_{\text{EN}}$, then the initial Conception representation of THEATER PLAY will have a positive score for the dimension corresponding to THEATRICAL WORK, but also for MATCH and PERFORM MUSIC, which are clearly unrelated to THEATER PLAY, as shown in Figure 1 (top left, dimensions related/unrelated to THEATER PLAY shown in green/red for illustrative purposes).

4.2 Cross-lingual concept disambiguation

The objective of the second step of Conception is, therefore, to refine the previously created vectors by excluding all the unsuitable components that were included due to lexical ambiguity. To do this, we exploit multilinguality so as to retain only those concept components that span across languages.

Consider again a concept c and its lexical vectors, one for each language. Then, for each lexical item w whose score is non-null in any lexical vector of c, we assume that the most relevant meaning c' of w with respect to c will appear the largest number of times across the different lexical vectors of c. More formally, we define the language span $\textsc{Span}(c'|c)$ of c' with respect to c as the number of languages where c' appears as the meaning of a word w with non-zero scores in the lexical vectors of c: $\textsc{Span}(c'|c) = |\{\, l \in L : \exists w \in V^l \wedge \mathbf{v}_c^l[w] \neq 0 \wedge c' \in \textsc{Senses}(w|l) \,\}|$, where V^l is the vocabulary of the words in language l, and $\textsc{Senses}(w|l)$ are the possible meanings of w in l. We can then build a new filtered vector $\hat{\mathbf{v}}_c$ where each component c' is zeroed if it is deemed unrelated to c:

$$\hat{\mathbf{v}}_c[c'] = \begin{cases} \mathbf{v}_c[c'] & \text{if } \exists w, l : c' = \operatorname*{argmax}_{c_w \in \textsc{Senses}(w|l)} \textsc{Span}(c_w|c) \\ 0 & \text{otherwise} \end{cases}$$

The resulting vector $\hat{\mathbf{v}}_c$ is a more accurate version of $\mathbf{v}_c$ in that some of its components have been zeroed based on the disambiguation of each word across languages. Following the previous example, the ambiguous word *tragedy*$_{\text{EN}}$ from the English lexical vector can be disambiguated thanks to *drame*$_{\text{FR}}$ from French. As shown in Figure 1 (top right), these two words share only the DRAMA meaning across languages, therefore Conception zeroes out all the other falsely positive components in the representation of THEATER PLAY.

4.3 Exploiting concept relations

After selecting the most important dimensions for each semantic vector, Conception takes advantage of the semantic relations defined in the BabelNet graph in order to directly inject explicit semantic knowledge into the representations. More formally, given a concept c and its vector $\hat{\mathbf{v}}_c$, for each concept c' corresponding to a component of $\hat{\mathbf{v}}_c$, Conception takes into account the value in $\hat{\mathbf{v}}_c$ of the neighboring concepts $c_n \in \mathrm{N}(c')$ of c' in the BabelNet graph:

$$\textsc{N-Score}(c'|c) = \sum_{c_n \in \mathrm{N}(c')} \frac{\hat{\mathbf{v}}_c[c_n]}{|\mathrm{N}(c')|}$$

Then, for each concept c, we create a new vector $\tilde{\mathbf{v}}_c$ from $\hat{\mathbf{v}}_c$ by adding the neighbors scored as above:

$$\tilde{\mathbf{v}}_c[c'] = \hat{\mathbf{v}}_c[c'] + \textsc{N-Score}(c'|c)$$

for each concept c', independently of whether $\hat{\mathbf{v}}_c[c']$ is 0 or not. As a result of this step, semantic information that was previously left unexpressed is made explicit, resulting in a new vector $\tilde{\mathbf{v}}_c$ where the number of non-zero dimensions is larger than in $\hat{\mathbf{v}}_c$. For instance, the representation of THEATER PLAY is now richer by including WORK OF ART as hypernym of THEATRICAL WORK, or ACT as meronym of DRAMA (see Figure 1, bottom left).

4.4 Symmetrizing concept representations

The previous step is "local" in that the injection of semantic knowledge into a concept representation does not depend on the representation of any other concept. In this final step, Conception enhances each concept representation with information contained in the representations of other concepts.

Let $G = (V, E)$ be a directed weighted graph where V is a set of concepts and E is a set of weighted relations between pairs of concepts. The vectors we have created so far can be seen as the weighted adjacency lists of each concept in the graph G. Given a concept $c \in V$ and its representation $\tilde{\mathbf{v}}_c$, if $\tilde{\mathbf{v}}_c[c'] > 0$, then there exists a relationship edge $e = (c, c') \in E$. We assume that, if e exists, then there should also exist an edge $\bar{e} = (c', c) \in E$, that is, $\tilde{\mathbf{v}}_{c'}[c]$ should be non-zero. If $\bar{e}$ already exists in $\tilde{\mathbf{v}}_{c'}$, we increase the weight of $\bar{e}$, otherwise we connect c' to c by creating $\bar{e}$ in the semantic vector of c'. In both cases, the updated weight of $\bar{e}$ depends on its previous weight (possibly null) and the importance of c with respect to c': $\bar{\mathbf{v}}_{c'}[c] = \tilde{\mathbf{v}}_{c'}[c] + f(c, c') \cdot \tilde{\mathbf{v}}_c[c']$, where $f(c, c')$ is the ratio between the weights of the two concepts c and c', $f(c, c') = \frac{\sigma(c)}{\sigma(c')}$, and $\sigma(c)$ computes the importance of a concept c over the whole BabelNet graph based on its weights in the vectors built as explained in Section 4.3:

$$\sigma(c) = \sum_{c'} \frac{\tilde{\mathbf{v}}_{c'}[c]}{\sum_{c''} \tilde{\mathbf{v}}_{c'}[c'']}$$

Getting back to our example, let the value of the GLOBE THEATRE dimension in the representation of THEATER PLAY be non-null. As shown in Figure 1 (bottom right), this step "connects" GLOBE THEATRE and THEATER PLAY by increasing the (possibly null) score of the THEATER PLAY dimension in the representation of the GLOBE THEATRE concept.

5 Semantic Word Similarity

We evaluate Conception on the Semantic Word Similarity task across 6 languages for a total of 6 multilingual and 10 cross-lingual datasets. Semantic Word Similarity is one of the most popular intrinsic benchmarks for the evaluation of representation techniques. Given two lexical items (words, multiword expressions or named entities), the task involves measuring their semantic closeness.

5.1 Experimental Setup

The evaluation of word-level representations in Semantic Word Similarity is often straightforward, since the semantic closeness of two lexical items can be measured directly by comparing the two corresponding representations. Instead, the application of concept-level representations like Conception's to word similarity requires consideration of all the possible senses of the two lexical items to be compared.

Word comparison. In the context of word sense and concept representations, the semantic distance between two words is traditionally computed as the similarity between their closest senses (Resnik, 1995; Budanitsky and Hirst, 2006). We use a variant of this comparison strategy so as to give more importance to the more frequent senses of a word:

$$\textsc{Sim}(w_1, w_2) = \max_{c_{w_1}, c_{w_2}} \frac{2 \cdot \textsc{Sim}(\mathbf{v}_{c_{w_1}}, \mathbf{v}_{c_{w_2}})}{\mathrm{R}_s(c_{w_1}, w_1) + \mathrm{R}_s(c_{w_2}, w_2)}$$

which computes the maximum similarity when considering all pairs of concepts c_{w_1} and c_{w_2} for words w_1 and w_2, where $\mathrm{R}_s(c_w, w)$ is the ranking of c_w among the senses of w sorted by decreasing value of

SemEval-2017	Best 4		All		C
Multilingual	r	ρ	r	ρ	%
SemEval baseline	.60	.60	.56	.56	–
NASARI$_{\text{lexical}}$	.71	.72	.69	.70	92
NASARI$_{\text{unified}}$	.70	.70	.69	.69	90
Conneau et al. (2017)	.59	.57	–	–	–
C. Numberbatch$_{\text{SE17}}$	.74	.75	.69	.70	–
C. Numberbatch$_{\text{19.08}}$	.74	.74	.70	.70	84
Conception$_{\text{concept selection}}$	.72	.73	.71	.71	92
Conception$_{\text{knowledge injection}}$	.74	.74	.72	.72	92
Conception$_{\text{symmetrization}}$	**.76**	**.77**	**.75**	**.75**	**95**

(a) Pearson (r) and Spearman (ρ) correlation performance in the multilingual Semantic Word Similarity task of SemEval-2017 (subtask 2.a). The *Best 4* score is the average over the 4 best results among the 5 languages, according to the official task evaluation. (C)overage indicates the percentage of word pairs covered by each approach across all the 6 languages.

SemEval-2017	Best 6		All		C
Cross-lingual	r	ρ	r	ρ	%
SemEval baseline	.60	.60	.56	.55	–
NASARI$_{\text{lexical}}$	.72	.73	.71	.71	93
NASARI$_{\text{unified}}$	.73	.73	.72	.72	91
Conneau et al. (2017)	.56	.53	–	–	–
C. Numberbatch$_{\text{SE17}}$	.75	.77	.69	.70	–
C. Numberbatch$_{\text{19.08}}$	.75	.76	.66	.67	83
XLM	–	–	.69	–	–
Conception	**.77**	**.78**	**.76**	**.77**	**95**

(b) Pearson (r) and Spearman (ρ) correlation performance in the cross-lingual Semantic Word Similarity task of SemEval-2017. The *Best 6* score is the average over the 6 best results among the 10 language pairs, as in the official SemEval-2017 evaluation. (C)overage indicates the percentage of word pairs covered by each approach across all the 10 language pairs.

Table 1: Pearson (r) and Spearman (ρ) correlation performance in the multilingual (left) and cross-lingual (right) Word Similarity tasks of SemEval-2017.

$\sigma(c_w)$ (see Section 4.4). We measure the semantic closeness of two senses (Sim in the above formula) using the square-rooted Absolute Weighted Overlap (Camacho-Collados et al., 2016) on their sparse vector representations.

Comparison systems. Vector representations for words, word senses and concepts can be split into two categories: sparse and dense representations. Conception, NASARI lexical and NASARI unified (see Sections 2 and 3) belong to the former category, so they are the most natural competitors in a comparison. However, over the last few years, dense vector representations have emerged as the most empirically effective type of representations in capturing syntactic and semantic relations between words. For this reason, we also compare Conception with the current state of the art in multilingual dense vector representation techniques. We include in the comparison multilingual word embeddings from the works of Conneau et al. (2017), created by aligning fastText embeddings in a unified space, Jawanpuria et al. (2019), obtained with language-specific transformations, and Speer et al. (2017, Conceptnet Numberbatch), built by retrofitting pre-trained word embeddings to the multilingual ConceptNet graph. We include both Conceptnet Numberbatch$_{19.08}$, which is latest version of the embeddings, and Conceptnet Numberbatch$_{\text{SE17}}$, which uses a complex strategy for out-of-vocabulary (OOV) words. To set a level playing field for all the systems, we assign the same score (0.5) for any OOV word pair.

5.2 Multilingual Word Similarity

Datasets. We evaluate Conception on SemEval-2017 Task 2.a (Camacho-Collados et al., 2017), a tough multilingual word similarity benchmark that provides hundreds of word pairs in 5 languages, namely English, Farsi, German, Italian and Spanish. SemEval-2017 is the ideal test bed for Conception since it features a low-resource language (Farsi) and, unlike other popular datasets such as SimLex-999 (Hill et al., 2015) and its translations, it also includes multiwords and named entities, which are difficult to model with word representations and are therefore often treated separately or ignored.

Results. Table 1a reports the average Pearson and Spearman correlation performance of Conception and all comparison systems on 5 languages (detailed per-language results are reported in the Appendix). Conception outperforms NASARI (both lexical and unified) – the current state of the art in sparse representations of concepts – by a remarkable margin (+5% across all languages). Our sparse vectors also outperform the state-of-the-art dense vectors of Conceptnet Numberbatch (+5% across all languages), while also providing considerably wider lexical coverage (+11% in Table 1a, last column). More generally, as long as BabelNet can provide a non-empty word-to-sense mapping, Conception can gracefully

borsa$_{\text{IT}}$: closest words in Conceptnet Numberbatch (CNNB)			*borsa*$_{\text{IT}}$: closest meanings in Conception	
IT		EN	HANDBAG	STOCK MARKET
borse$_{\text{IT}}$	(bags)	*evening bag*$_{\text{EN}}$	PURSE	MARKETPLACE (economics)
borsaio$_{\text{IT}}$	(bag maker)	*luggage store*$_{\text{EN}}$	SHOPPING BAG	FINANCIAL ASSET
borsata$_{\text{IT}}$	(bagful)	*handbag*$_{\text{EN}}$	BAG	SHARE (of capital stock)
borsello$_{\text{IT}}$	(purse)	*satchel*$_{\text{EN}}$	CONTAINER	PRICE

Table 2: Left: the Italian and English words nearest to *borsa*$_{\text{IT}}$ in CNNB according to cosine similarity (manual translations of the Italian words in brackets). Notice that no finance-related word appears among its nearest neighbors. Right: Conception models the HANDBAG and the STOCK MARKET meanings of *borsa*$_{\text{IT}}$ independently: the representations are distinct and non-overlapping, as evidenced by their top-scoring concept dimensions.

scale across 284 languages.

5.3 Cross-lingual Word Similarity

Datasets. We also compare Conception on the SemEval-2017 Task 2.b (Camacho-Collados et al., 2017, Cross-lingual Semantic Word Similarity). This task is similar to the multilingual Word Similarity task described in Section 5.2, with the key difference that the two lexical items to compare belong to different languages. SemEval-2017 includes 10 cross-lingual datasets from 5 languages, namely English, German, Spanish, Italian and Farsi. Each dataset contains around 1,000 entries that compare words, multiword expressions and named entities across the aforementioned languages.

Results. Conception provides a notable increase in correlation performance over the state-of-the-art sparse vector representations of NASARI (both lexical and unified) across all the cross-lingual datasets of the task, averaging a 5% and a 6% absolute improvement in Pearson and Spearman correlations respectively, as shown in Table 1b (see the Appendix for consistently state-of-the-art pairwise figures).

Furthermore, our sparse vector representations outperform the state-of-the-art dense vector representations of Conceptnet Numberbatch (CNNB), Conneau et al. (2017) and Jawanpuria et al. (2019) in every language pair in the task except for the English-German test, where the results are comparable with CNNB. The difference in performance is remarkable in the evaluations that involve Farsi, and this once again highlights the robustness of Conception on low-resource languages. In Table 1b, we also report the score of XLM (Conneau and Lample, 2019), a language model trained with an explicit cross-lingual objective.

5.4 Analysis

Ablation study. In order to better appreciate the contribution of each step of the Conception algorithm to the final results, we analyzed the difference in performance between NASARI and the representations created after a) selecting the concepts through cross-lingual disambiguation (Section 4.2), b) exploiting the semantic relations in the BabelNet graph to inject knowledge (Section 4.3), and c) symmetrizing concept relations (Section 4.4), i.e., the vectors produced at the end of the Conception algorithm. Table 1a (bottom) reports a comparison of the results of the above representations in the multilingual word similarity task of SemEval-2017. Each step of the Conception algorithm incrementally improves the performance of the representations of the previous step, which is particularly evident for the last step. We observed comparable improvements in the cross-lingual setting.

Case study. We conducted an in-depth analysis of the correlation performance obtained by Conception, CNNB, and the NASARI lexical vectors in the SemEval-2017 Italian benchmark (subtask 2.a), with the aim of understanding where Conception makes the difference. Conception shines in capturing semantic similarity between word pairs with a high gold similarity score (≥ 0.75). In contrast to Conception, for example, CNNB struggles in capturing semantic similarity in synonym pairs ($\text{sim}_{\text{gold}} = 1.0$) which involve highly-ambiguous words such as *schermo*$_{\text{IT}}$ – *monitor*$_{\text{IT}}$ ($\text{sim}_{\text{pred}} < 0.5$), or multiword expressions such as *sclerosi multipla*$_{\text{IT}}$ – *sclerosi a placche*$_{\text{IT}}$ ($\text{sim}_{\text{pred}} = 0.5$), or named entities such as

						Concatenation of ALL datasets				
	SE2	**SE3**	**SE07**	**SE13**	**SE15**	**Nouns**	**Verbs**	**Adj**	**Adv**	**ALL**
Raganato et al. (2017a)	72.0	69.1	64.8	66.9	71.5	71.5	57.5	75.0	83.8	69.9
BERT$_{\text{Large}}$	76.3	73.2	66.2	71.7	74.1	–	–	–	–	73.5
KnowBERT (Peters et al., 2019)	–	–	–	–	–	–	–	–	–	75.1
SensEmBERT (Scarlini et al., 2020a)	–	–	–	74.8	–	–	–	–	–	–
LMMS (Loureiro and Jorge, 2019)	76.3	75.6	68.1	75.1	77.0	78.0	64.0	80.7	84.5	75.4
Conception $_{\text{SenseEmBERT}}$	–	–	–	75.9	–	–	–	–	–	–
Conception $_{\text{LMMS}}$	**77.1**	**76.4**	**70.3**	**76.2**	**77.2**	**78.7**	**65.6**	**81.1**	**84.7**	**76.4**

Table 3: WSD results in F$_1$ scores on Senseval-2 (SE2), Senseval-3 (SE3), SemEval-2007 (SE07), SemEval-2013 (SE13), SemEval-2015 (SE15), and the concatenation of all the datasets (ALL).

DeepMind$_{\text{IT}}$ – *Google DeepMind*$_{\text{IT}}$ ($\text{sim}_{\text{pred}} = 0.5$). Another representative example is the synonym pair *borsa*$_{\text{IT}}$ – *mercato azionario*$_{\text{IT}}$, where the first term means either HANDBAG or STOCK MARKET, and the second term assumes the latter meaning only. In Table 2 (right), we show how the two concepts of HANDBAG and STOCK MARKET are modelled separately by Conception based on their closest meanings. The STOCK MARKET meaning of *borsa*$_{\text{IT}}$ enables Conception to capture the synonymity *borsa*$_{\text{IT}}$ – *mercato azionario*$_{\text{IT}}$ ($\text{sim}_{\text{pred}} = \text{sim}_{\text{gold}} = 1.0$). In contrast, CNNB fails to do so ($\text{sim}_{\text{pred}} = 0.42$): the reason is that, in a similar vein to other word-level representations, CNNB conflates the various meanings of an ambiguous word into a single vector where the predominant meaning may overshadow the other meanings. In our example, the STOCK MARKET concept is overshadowed by the HANDBAG concept in the word representation of *borsa*$_{\text{IT}}$ of CNNB (Table 2, left). Conversely, Conception models the two concepts independently (Table 2, right), and it is consequently able to correctly capture similarity in more difficult settings.

6 Word Sense Disambiguation

Word Sense Disambiguation (WSD) – the task of assigning the correct meaning to a target word in a context – is considered to be a fundamental step towards natural language understanding (Navigli, 2018). As with many other tasks, WSD has benefited greatly from the recent advances in other fields, such as language modelling (Scarlini et al., 2020b), game theory (Tripodi and Navigli, 2019), structured knowledge integration (Bevilacqua and Navigli, 2020), definition modelling (Bevilacqua et al., 2020) and label propagation (Barba et al., 2020; Pasini and Navigli, 2020), *inter alia*. Our experiments show that Conception can be used to create state-of-the-art sense embeddings, demonstrating empirically that our approach provides high-quality knowledge that is still not captured by recent language models.

Experimental setup. We start from state-of-the-art, precomputed sense embeddings and adopt a simple strategy to enrich such representations with Conception in order to evaluate its effectiveness in WSD. First, we create an embedding $\mathbf{e}_c$ for a concept c by averaging the precomputed embeddings $\mathbf{e}_s$ of each word sense s that can be used to express c: $\mathbf{e}_c = \sum_{s \in c} \frac{\mathbf{e}_s}{|\{s \in c\}|}$. Then, given a word sense s and its corresponding concept c, we build a new word sense embedding $\mathbf{e}'_s$ by adding to $\mathbf{e}_s$ each concept embedding $\mathbf{e}_{c'}$ weighted by the ranking of concept c' in the Conception representation of c:

$$\mathbf{e}'_s = \alpha \cdot \mathbf{e}_s + (1 - \alpha) \cdot \frac{\sum_{c'} \mathbf{e}_{c'} \cdot \text{R}(c', \mathbf{v}_c)^{-1}}{\sum_{c'} \text{R}(c', \mathbf{v}_c)^{-1}} \quad (1)$$

where $\alpha = 0.5$, and $\text{R}(c', \mathbf{v}_c)$ is the ranking of concept c' in the Conception representation of c.

Comparison systems. We consider the following state-of-the-art sense embeddings: LMMS (Loureiro and Jorge, 2019), a supervised technique that combines BERT contextualized embeddings with knowledge from WordNet; SensEmBERT (Scarlini et al., 2020a), a knowledge-based BERT-based approach that enriches contextualized embeddings with knowledge from Wikipedia, BabelNet and NASARI vectors; BERT$_{\text{Large}}$ as reported by Loureiro and Jorge (2019). We compare the above embeddings against those obtained when enriching the LMMS and SensEmBERT embeddings with Conception based on Eq. 1 (Conception$_{\text{LMMS}}$ and Conception$_{\text{SensEmBERT}}$ hereafter).

For all embeddings, WSD is performed as customary in the literature: given a word w in context, we choose the sense s of w whose vector is closest to the contextual BERT representation of w according to cosine similarity. We also include KnowBERT (Peters et al., 2019), a language model which exploits multiple knowledge bases.

Datasets. We empirically assess our sense embeddings on the unified evaluation framework for English WSD proposed by Raganato et al. (2017b), which comprises Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 (Snyder and Palmer, 2004), SemEval-2007 (Pradhan et al., 2007), SemEval-2013 (Navigli et al., 2013), and SemEval-2015 (Moro and Navigli, 2015).

Results. Table 3 shows that, on the concatenation of all the English WSD datasets, Conception$_{\text{LMMS}}$ surpasses the state-of-the-art supervised representations of LMMS (+1.0% F$_1$), while Conception$_{\text{SensEmBERT}}$ outperforms the state-of-the-art knowledge-based representations of SensEmBERT (+1.1% F$_1$). In particular, we would like to highlight three main findings: i) Conception encodes a non-trivial amount of knowledge that is not included in LMMS, even though this latter already exploits BERT and the WordNet semantic graph; ii) Conception builds its representations from NASARI, which only covers nominal concepts. Nevertheless, it also produces robust representations for verbal concepts (+1.6% in F$_1$ score over LMMS); iii) while Conception and SensEmBERT are both knowledge-based techniques relying on BabelNet, the former is still able to inject meaningful knowledge into the latter. In general, the knowledge included by Conception is orthogonal to existing sense representations. While we have here opted to enrich existing sense embeddings with a simple yet effective technique, we envisage that more sophisticated uses of Conception can lead to further improvements in WSD.

7 Conclusion

In this paper we presented Conception, a novel knowledge-based technique for modelling concepts and named entities. Its key innovation lies in setting multilinguality as the cornerstone of the learning process to build language-agnostic and human-readable concept vector representations.

Evaluated across multiple multilingual and cross-lingual Semantic Word Similarity datasets, Conception shows state-of-the-art results not only compared to concept representations such as NASARI, but also to multilingual word embeddings such as Conceptnet Numberbatch and cross-lingual language models such as XLM. Additionally, our concept representations are particularly robust on resource-poor languages, like Farsi, along the lines of recent work in Semantic Parsing and Semantic Role Labeling aimed at bridging the gap between languages (Blloshmi et al., 2020; Conia and Navigli, 2020). Finally, Conception can be seamlessly applied to a downstream task: in Word Sense Disambiguation, it improves over state-of-the-art supervised and knowledge-based sense embeddings, showing that Conception encodes information that is still not captured by BERT-based contextualized representations.

Furthermore, our approach produces much more than concept representations: since each concept is described by the relationships it has with other concepts, Conception can be seen as a weighted directed graph where each node is a concept whose vector representation is also its weighted adjacency list. This paves the way to a whole set of possible applications for Conception: from graph-based concept embeddings to semantics-first sentence and document representations. The complete set of concept vectors is available at `https://github.com/SapienzaNLP/conception`.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 under the European Union's Horizon 2020 research and innovation programme.

This work was supported in part by the MIUR under grant "Dipartimenti di eccellenza 2018-2022" of the Department of Computer Science of Sapienza University.

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In Julia Hockenmaier and Sebastian Riedel, editors, *Proceedings of the Seventeenth Conference on Computational Natural Language Learning, CoNLL 2013, Sofia, Bulgaria, August 8-9, 2013*, pages 183–192.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *CoRR*, abs/1602.01925.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 789–798.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Edoardo Barba, Luigi Procopio, Niccolo Campolungo, Tommaso Pasini, and Roberto Navigli. 2020. MuLaN: Multilingual label propagation for Word Sense Disambiguation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3837–3844.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in Word Sense Disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864, July.

Michele Bevilacqua, Marco Maru, and Roberto Navigli. 2020. Generationary or: "How we went beyond sense inventories and learned to gloss". In *The 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, November.

Rexhina Blloshmi, Rocco Tripodi, and Roberto Navigli. 2020. XL-AMR: Enabling Cross-Lingual AMR parsing with transfer learning techniques. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguistics*, 5:135–146.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 1352–1362.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

José Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *J. Artif. Intell. Res.*, 63:743–788.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artif. Intell.*, 240:36–64.

José Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. SemEval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel M. Cer, and David Jurgens, editors, *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*, pages 15–26.

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 261–270.

Simone Conia and Roberto Navigli. 2020. Bridging the gap in multilingual Semantic Role Labeling: A language-agnostic approach. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020*.

Simone Conia, Fabrizio Brignone, Davide Zanfardino, and Roberto Navigli. 2020. InVeRo: Making Semantic Role Labeling accessible with intelligible verbs and roles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 7057–7067.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *CoRR*, abs/1710.04087.

Steven Derby, Paul Miller, Brian Murphy, and Barry Devereux. 2018. Using sparse semantic embeddings learned from multimodal text and image data to model human conceptual knowledge. In Anna Korhonen and Ivan Titov, editors, *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, pages 260–270.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Philip Edmonds and Scott Cotton. 2001. SENSEVAL-2: overview. In Judita Preiss and David Yarowsky, editors, *Proceedings of Second International Workshop on Evaluating Word Sense Disambiguation Systems, SENSEVAL@ACL 2001, Toulouse, France, July 5-6, 2001*, pages 1–5.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In Gosse Bouma and Yannick Parmentier, editors, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 462–471.

Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah A. Smith. 2015. Sparse overcomplete 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 of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1491–1500.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 473–483.

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

Ignacio Iacobacci and Roberto Navigli. 2019. LSTMEmbed: Learning word and sense representations from a large semantically annotated corpus with long short-term memories. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1685–1695.

Ignacio Iacobacci, 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 of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 95–105.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Pratik Jawanpuria, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra. 2019. Learning multilingual word embeddings in latent metric space: A geometric approach. *Trans. Assoc. Comput. Linguistics*, 7:107–120.

Tom Kenter and Maarten de Rijke. 2015. Short text similarity with word embeddings. In James Bailey, Alistair Moffat, Charu C. Aggarwal, Maarten de Rijke, Ravi Kumar, Vanessa Murdock, Timos K. Sellis, and Jeffrey Xu Yu, editors, *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, pages 1411–1420.

Aykut Koç, Ihsan Utlu, Lutfi Kerem Senel, and Haldun M. Özaktas. 2018. Imparting interpretability to word embeddings. *CoRR*, abs/1807.07279.

Pierre Lafon. 1980. Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique*, 1(1):127–165.

Yoav Levine, Barak Lenz, Or Dagan, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2019. SenseBERT: Driving some sense into BERT. *CoRR*, abs/1908.05646.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 302–308.

Jiwei Li and Dan Jurafsky. 2015. Do multi-sense embeddings improve natural language understanding? In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1722–1732.

Daniel Loureiro and Alípio Jorge. 2019. Language modelling makes sense: Propagating representations through wordnet for full-coverage word sense disambiguation. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5682–5691.

Diego Marcheggiani, Anton Frolov, and Ivan Titov. 2017. A simple and accurate syntax-agnostic neural model for dependency-based semantic role labeling. In Roger Levy and Lucia Specia, editors, *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017*, pages 411–420.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In Yoav Goldberg and Stefan Riezler, editors, *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 51–61.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In Daniel M. Cer, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pages 288–297.

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

Roberto Navigli and Simone Paolo Ponzetto. 2012b. Joining forces pays off: Multilingual joint word sense disambiguation. In Jun'ichi Tsujii, James Henderson, and Marius Pasca, editors, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, pages 1399–1410.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 task 12: Multilingual word sense disambiguation. In Mona T. Diab, Timothy Baldwin, and Marco Baroni, editors, *Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2013, Atlanta, Georgia, USA, June 14-15, 2013*, pages 222–231.

Roberto Navigli. 2018. Natural Language Understanding: Instructions for (Present and Future) Use. In *Proc. of the 27th International Joint Conference on Artificial Intelligence (IJCAI-18)*, pages 5697–5702, Stockholm, Sweden.

Hien T. Nguyen, Phuc H. Duong, and Erik Cambria. 2019. Learning short-text semantic similarity with word embeddings and external knowledge sources. *Knowl. Based Syst.*, 182.

Sungjoon Park, JinYeong Bak, and Alice Oh. 2017. Rotated word vector representations and their interpretability. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 401–411.

Tommaso Pasini and Roberto Navigli. 2020. Train-O-Matic: Supervised word sense disambiguation with no (manual) effort. *Artif. Intell.*, 279.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237.

Matthew E. Peters, Mark Neumann, Robert L. Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 43–54.

Mohammad Taher Pilehvar and José Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1267–1273.

Mohammad Taher Pilehvar, José Camacho-Collados, Roberto Navigli, and Nigel Collier. 2017. Towards a seamless integration of word senses into downstream NLP applications. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1857–1869.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task-17: English lexical sample, SRL and all words. In Eneko Agirre, Lluís Màrquez i Villodre, and Richard Wicentowski, editors, *Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval@ACL 2007, Prague, Czech Republic, June 23-24, 2007*, pages 87–92.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017a. Neural sequence learning models for word sense disambiguation. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1156–1167.

Alessandro Raganato, José Camacho-Collados, and Roberto Navigli. 2017b. Word sense disambiguation: A unified evaluation framework and empirical comparison. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 99–110.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95, Montréal Québec, Canada, August 20-25 1995, 2 Volumes*, pages 448–453. Morgan Kaufmann.

Sascha Rothe and Hinrich Schütze. 2016. Word embedding calculus in meaningful ultradense subspaces. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020a. SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8758–8765.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020b. With More Contexts Comes Better Performance: Contextualized Sense Embeddings for All-Round Word Sense Disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*.

Lutfi Kerem Senel, Ihsan Utlu, Veysel Yücesoy, Aykut Koc, and Tolga Çukur. 2018. Semantic structure and interpretability of word embeddings. *IEEE ACM Trans. Audio Speech Lang. Process.*, 26(10):1769–1779.

Karan Singhal, Karthik Raman, and Balder ten Cate. 2019. Learning multilingual word embeddings using image-text data. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 68–77.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Benjamin Snyder and Martha Palmer. 2004. The english all-words task. In *Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, SENSEVAL@ACL 2004, Barcelona, Spain, July 25-26, 2004*.

Robyn Speer and Joanna Lowry-Duda. 2017. Conceptnet at semEval-2017 task 2: Extending word embeddings with multilingual relational knowledge. In Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel M. Cer, and David Jurgens, editors, *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*, pages 85–89.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In Satinder P. Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451.

Rocco Tripodi and Roberto Navigli. 2019. Game theory meets embeddings: a unified framework for Word Sense Disambiguation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 88–99, Hong Kong, China, November.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. Allennlp interpret: A framework for explaining predictions of NLP models. In Sebastian Padó and Ruihong Huang, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019 - System Demonstrations*, pages 7–12.

Guangyou Zhou, Tingting He, Jun Zhao, and Po Hu. 2015. Learning continuous word embedding with metadata for question retrieval in community question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 250–259.

A Multilingual Word Similarity

Table 4 includes the complete results over all the 5 languages in the Multilingual Word Similarity subtask of SemEval-2017. As can be seen, Conception shows a much less abrupt decrease in performance when evaluated on Farsi compared to other knowledge-based, purely distributional and knowledge-enhanced approaches such as NASARI (Camacho-Collados et al., 2016), GeoMM (Jawanpuria et al., 2019), and Conceptnet Numberbatch (Speer and Lowry-Duda, 2017), respectively.

SemEval-2017	**English**		**German**		**Spanish**		**Italian**		**Farsi**	
Multilingual	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
SemEval baseline*	.68	.68	.51	.51	.60	.60	.60	.59	.41	.40
NASARI $_{\text{lexical}}$	.75	.75	.64	.64	.73	.74	.72	.73	.62	.62
NASARI $_{\text{unified}}$	.74	.73	.63	.63	.73	.73	.72	.72	.62	.63
Conneau et al. (2017)	.59	.57	.60	.58	.59	.57	.56	.54	–	–
Jawanpuria et al. (2019)	.59	.56	.62	.61	.53	.52	–	–	–	–
Conceptnet Numberbatch $_{\text{SE17}}$ *	.78	**.80**	.70	.70	.73	.75	.73	.75	.51	.50
Conceptnet Numberbatch $_{19.08}$	**.79**	**.80**	**.75**	**.75**	.71	.71	.70	.70	.56	.54
Conception $_{\text{concept selection}}$	.77	.77	.65	.65	.74	.74	.73	.74	.66	.66
Conception $_{\text{knowledge injection}}$	.78	.79	.66	.67	.75	.76	.74	.75	.67	.67
Conception $_{\text{concept symmetrization}}$	**.79**	.79	.70	.71	**.77**	**.79**	**.77**	**.79**	**.69**	**.68**

Table 4: Pearson (r) and Spearman (ρ) correlation performance of Conception compared with the current state of the art in the SemEval-2017 multilingual Semantic Word Similarity task (subtask 2.a). The scores of the SemEval-2017 baseline and Conceptnet Numberbatch$_{\text{SE17}}$ are taken directly from Camacho-Collados et al. (2017) and Speer and Lowry-Duda (2017), respectively.

B Cross-lingual Word Similarity

Table 5 includes the complete results over all the 10 language pairs in the Cross-lingual Word Similarity subtask of SemEval-2017. Conception achieves a new state of the art in 9 of the 10 language pairs, and, once again, it is particularly robust in language pairs where Farsi, a low-resource language, is involved.

SemEval-2017	**DE-ES**		**DE-FA**		**DE-IT**		**EN-DE**		**EN-ES**	
Cross-lingual	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
SemEval baseline*	.55	.55	.46	.45	.56	.56	.60	.59	.64	.63
NASARI $_{\text{lexical}}$	.70	.70	.67	.67	.70	.70	.71	.71	.74	.75
NASARI $_{\text{unified}}$	.71	.71	.68	.68	.72	.72	.73	.73	.76	.77
Conneau et al. (2017)	.56	.54	–	–	.53	.51	.57	.54	.56	.54
Jawanpuria et al. (2019)	.54	.52	–	–	–	–	.58	54	.51	.49
Conceptnet Numberbatch $_{\text{SE17}}$ *	.72	.74	.59	.59	.74	.75	.76	**.77**	.75	.77
Conceptnet Numberbatch $_{\text{19.08}}$	.73	.74	.66	.66	.74	.74	**.77**	**.77**	.74	.75
Conception	**.75**	**.76**	**.74**	**.74**	**.76**	**.77**	**.77**	**.77**	**.79**	**.80**

SemEval-2017	**EN-FA**		**EN-IT**		**ES-FA**		**ES-IT**		**IT-FA**	
Cross-lingual	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
SemEval baseline*	.52	.49	.65	.65	.49	.47	.60	.59	.50	.48
NASARI $_{\text{lexical}}$	.69	.70	.75	.76	.69	.70	.73	.74	.68	.69
NASARI $_{\text{unified}}$	.72	.72	.76	.77	.71	.72	.75	.75	.70	.71
Conneau et al. (2017)	–	–	.56	.52	–	–	.56	.54	–	–
Jawanpuria et al. (2019)	–	–	–	–	–	–	–	–	–	–
Conceptnet Numberbatch $_{\text{SE17}}$ *	.60	.59	.77	.79	.62	.63	.74	.77	.60	.61
Conceptnet Numberbatch $_{\text{19.08}}$	.67	.67	.75	.76	.65	.65	.71	.72	.65	.65
Conception	**.74**	**.75**	**.79**	**.80**	**.75**	**.75**	**.78**	**.80**	**.73**	**.74**

Table 5: Pearson (r) and Spearman (ρ) correlation performance of Conception compared with the current state of the art in the cross-lingual Semantic Word Similarity task of SemEval-2017 (subtask 2.b). The scores of the SemEval-2017 baseline and Conceptnet Numberbatch$_{\text{SE17}}$ are taken directly from Camacho-Collados et al. (2017) and Speer and Lowry-Duda (2017), respectively.