Building Lexical Vector Representations from Concept Definitions

Danilo S. Carvalho and **Minh Le Nguyen**
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi City, Ishikawa, Japan
{danilo, nguyenml}@jaist.ac.jp

Abstract

The use of distributional language representations have opened new paths in solving a variety of NLP problems. However, alternative approaches can take advantage of information unavailable through pure statistical means. This paper presents a method for building vector representations from meaning unit blocks called concept definitions, which are obtained by extracting information from a curated linguistic resource (Wiktionary). The representations obtained in this way can be compared through conventional cosine similarity and are also interpretable by humans. Evaluation was conducted in semantic similarity and relatedness test sets, with results indicating a performance comparable to other methods based on single linguistic resource extraction. The results also indicate noticeable performance gains when combining distributional similarity scores with the ones obtained using this approach. Additionally, a discussion on the proposed method's shortcomings is provided in the analysis of error cases.

1 Introduction

Vector-based language representation schemes have gained large popularity in Natural Language Processing (NLP) research in the recent years. Their success comes from both the asserted benefits in several NLP tasks and from the ability to built them from unannotated textual data, widely available in the World Wide Web. The tasks benefiting from vector representations include Part-of-Speech (POS) tagging (dos Santos and Zadrozny, 2014), dependency parsing (Bansal et al., 2014), Named Entity Recognition (NER) (Seok et al., 2016), Machine Translation (Sutskever et al., 2014), among others.

Such representation schemes are, however, not an all-in-one solution for the many NLP application scenarios. Thus, different representation methods were developed, each one focusing in a limited set of concerns, e.g., semantic relatedness measurement (Mikolov et al., 2013; Pennington et al., 2014) and grammatical dependencies (Levy and Goldberg, 2014). Most of the popular methods are based on a *distributional* approach: the meaning of a word is defined by the context of its use, i.e., the neighboring words. However, distributional representations carry no explicit linguistic information and cannot easily represent some important semantic relationships, such as synonymy and antonymy (Nguyen et al., 2016). Further problems include the difficulty in obtaining representations for out-of-vocabulary (OOV) words and complex constructs (collocations, idiomatic expressions), the lack of interpretable representations (Faruqui and Dyer, 2015), and the necessity of specific model construction for cross-language representation.

This paper presents a linguistically motivated language representation method, aimed at capturing and providing information unavailable on distributional approaches. Our contributions are: (i) a technique for building conceptual representations of linguistic elements (morphemes, words, collocations, idiomatic expressions) from a single collaborative language resource (*Wiktionary* [1]); (ii) a method of combining said representations and comparing them to obtain a semantic similarity measurement. The conceptual representations, called *Term Definition Vectors*, address more specifically the issues of semantic relationship analysis, out-of-vocabulary word interpreta-

[1] www.wiktionary.org

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 905–915,
Valencia, Spain, April 3-7, 2017. ©2017 Association for Computational Linguistics

tion and cross-language conceptual mapping. Additionally, they have the advantages of being interpretable by humans and easy to operate, due to their sparsity. Experiments were conducted with the *SimLex-999* (Hill et al., 2015) test collection for word similarity, indicating a good performance in this task and exceeding the performance of other single information source studies, when combined with a distributional representation and Machine Learning. Error analysis was also conducted to understand the strengths and weaknesses of the proposed method.

The remainder of this paper is organized as follows: Section 2 presents relevant related works and highlights their similarities and differences to this research. Section 3 explains our approach in detail, covering its linguistic motivation and the characteristics of both representation model and comparison method. Section 4 describes the experimental evaluation and discusses the evaluation results and error analysis. Finally, Section 5 offers a summary of the findings and some concluding remarks.

2 Related Work

In order to address the limitations of the most popular representation schemes, approaches for all-in-one representation models were also developed (Pilehvar and Navigli, 2015; Derrac and Schockaert, 2015). They comprise a combination of techniques applied over different data sources for different tasks. Pilehvar and Navigli (2015) presented a method for combining Wiktionary and Wordnet (Fellbaum and others, 1998) sense information into a semantic network and a corresponding relatedness similarity measurement. The method is called ADW (Align, Disambiguate, Walk), and works by first using a Personalized PageRank (PPR) (Haveliwala, 2002) algorithm for performing a random walk on the semantic network and compute a *semantic signature* of a linguistic item (sense, word or text): a probability distribution over all entities in the network where the weights are estimated on the basis of the network's structural properties. Two linguistic items are then aligned and disambiguated by finding their two closest senses, comparing their semantic signatures under a set of vector and rank-based similarity measures (JensenShannon divergence, cosine, Rank-Biased Overlap, and Weighted Overlap). ADW achieved state-of-the-art performance in several semantic relatedness test sets, covering words, senses and entire texts.

Recski et al. (2016) presented a hybrid system for measuring the semantic similarity of word pairs, using a combination of four distributional representations (*SENNA* (Collobert and Weston, 2008), (Huang et al., 2012), *word2vec* (Mikolov et al., 2013), and *GloVe* (Pennington et al., 2014)), *WordNet*-based features and *4lang* (Kornai, 2010) graph-based features to train a RBF kernel Support Vector Regression on the SimLex-999 (Hill et al., 2015) data set. This system achieved state-of-the-art performance in SimLex-999.

The work presented in this paper takes a similar approach to Pilehvar and Navigli (2015), but stops short on obtaining a far reaching concept graph. Instead, it focuses on exploring the details of each sense definition. This includes term etymologies, morphological decomposition and translation links, available in Wiktionary. Another difference is that the translation links are used to map senses between languages in this work, whereas they are used for bridging gaps between sense sets on monolingual text in Pilehvar and Navigli (2015).

Another concern regarding distributional representations is their lack of interpretability from a linguistic standpoint. Faruqui and Dyer (2015) addresses this point, relying on linguistic information from Wordnet, Framenet (F. Baker et al., 1998), among other sources (excluding Wiktionary), to build interpretable word vectors. Such vectors accommodate several types of information, ranging from Part-of-Speech (POS) tags to sentiment classification and polarity. The obtained linguistic vectors achieved very good performance in a semantic similarity test. Those vectors, however, do not include morphological and translation information, offering discrete, binary features.

Regarding the extraction of definition data from Wiktionary, an effective approach is presented by Zesch et al. (2008a), which is also used for building a semantic representation (Zesch et al., 2008b). However, the level of detail and structure format obtained by such method was not deemed adequate for this work and an alternative extraction method was developed (Sections 3.2 and 3.3).

3 Term Definition Vectors

The basic motivation for the representation model here described is both linguistic and epistemic:

trying to represent knowledge as a set of individual concepts that relate to one another and are related to a set of terms. This idea is closely related to the Ogden/Richards *triangle of reference* (Ogden et al., 1923) (Figure 1), which describes a relationship between linguistic symbols and the objects they represent. The following notions are then defined:

- *Concept*: The unit of knowledge. Represents an individual meaning, e.g., rain (as in the natural phenomenon), and can be encoded into a term (symbol). It corresponds to the "thought or reference" from the triangle of reference.

- *Term*: A unit of perception. In text, it can be mapped to fragments ranging from morphemes to phrases. Each one can be decoded into one or more *concepts*. Stands for the "symbol" in the triangle of reference.

- *Definition*: A minimal, but complete explicitation of a concept. It comprises the textual explanation of the concept (sense) and its links to other concepts in a knowledge base, corresponding to the "symbolizes" relationship in the triangle of reference. The simplest case is a dictionary definition, consisting solely of a short explanation (typically a single sentence), with optional term highlights, linking to other dictionary entries. The information used for building definitions in this work is described in Section 3.3.

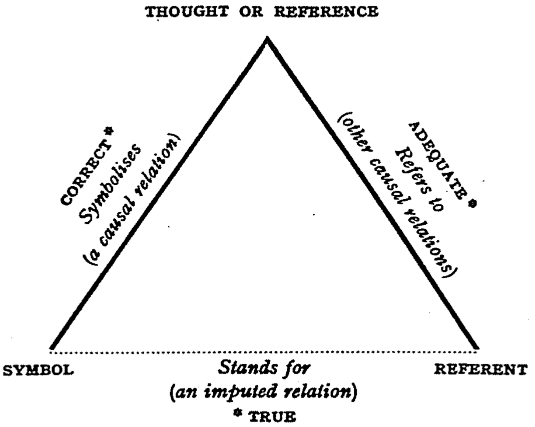


Figure 1: Ogden/Richards triangle of reference, also known as *semiotic triangle*. Describes a relationship between linguistic symbols and the objects they represent. (Ogden et al., 1923)

3.1 Distributional & Definitional Semantics

Distributional approaches for language representation, also known as *latent* or *statistical* semantics, are rooted in what is called the *distributional hypothesis* (Sahlgren, 2008). This concept stems from the notion that words are always used in a context, and it is the context that defines their meaning. Thus, the meaning of a term is concealed, i.e. latent, and can be revealed by looking at its context. In this sense, it is possible to define the meaning of a term to be a function of its neighboring term frequencies (co-occurrence). Using different definitions for "neighbor", e.g., adjacent words in *word2vec* (Mikolov et al., 2013) and "modifiers in a dependency tree" (Levy and Goldberg, 2014), it is possible to produce a variety of vector spaces, called *embeddings*. Good embeddings enable the use of vector operations on words, such as comparison by cosine similarity. They also solve the data sparsity problem of large vocabularies, working as a dimensionality reduction method. There are, however, semantic elements that are not directly related to context, and thus are not well represented by distributional methods, e.g., the antonymy and hypernymy relations. Furthermore, polysemy can bring potential ambiguity problems in cases where the vectors are only indexed by surface form (word → embedding).

An alternative line of thinking is to define the meanings first and then associate the corresponding terms (reference → symbol). In this notion, meanings are *explicit* and need only to be resolved, i.e., disambiguated, for any given term. Concepts are thus represented by prior definitions instead of distributions over corpora, hence the name "*definitional semantics*" is used in this work to generalize such approaches.

To illustrate the difference between both approaches, a simple analogy can be made, where a person reads a book with difficult or new vocabulary. The distributional approach would be akin to reading the book while trying to guess the meaning of the unknown words by context. If the book is long, as the reading progresses, the guesses tend to become more accurate, as a human will try to piece together the information patterns surrounding the new words. On the other hand, the definitional approach would be equivalent to reading the entire contents of a dictionary before reading the book. The main advantage of the former is independence from any previously compiled knowledge base, e.g, a dictionary, which are subject to completeness and correctness concerns. The latter offers answers for the rarer words that are difficult to guess and the possibility to explain exactly how a certain meaning was inferred (interpretability).

The proposed definitional representation is then obtained through the following strategy:

1. Formalization of the basic unit of knowledge: the concept.

2. Information extraction from a linguistic resource into a set of concepts.

3. Lexical association: term ↔ concept.

4. Definition of a term as a composition (mixture) of concepts, allowing partial or complete disambiguation.

Figure 2 illustrates the process. A term is said ambiguous if it is composed by more than one concept. Therefore in this context, disambiguation is the action of reducing the number of concepts in a term's composition. This can be done by collecting additional information about the term, such as Part-of-Speech. A complete disambiguation reduces the composition to a single concept.

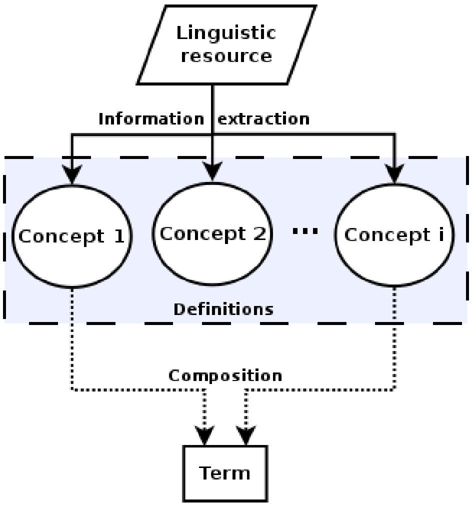


Figure 2: Process of definitional representation. Given a set of concepts obtained from a linguistic resource, a term can be defined as a composition of concepts. A term is said ambiguous if it is composed by more than one concept.

3.2 Linguistic Information Extraction

Wiktionary [2] was used as the single linguistic resource. Wiktionary is a collaborative lexical resource, comprising millions of vocabulary entries from several languages. It includes contextual information, etymology, semantic relations, translations, inflections, among other types of information for each entry. Its contents are actively curated by a large, global community. This choice was motivated by a several reasons, more importantly:

- It is the largest lexical resource openly available for the public, covering more than 10 million lexical entries from 172 languages.

- It is constantly updated. Daily changes are consolidated in monthly releases.

- Entries are organized in a way that separates each meaning of a term, simplifying definition extraction.

- Entries include range from morphemes, e.g., "pre-", to idiomatic expressions, e.g., "take matters into one's own hands".

The data available from Wiktionary is semi-structured, composed of a set of markup documents, one for each entry, following a reasonably consistent standard of annotations for each language covered. In order to extract the linguistic information, an application was developed to convert the markup into a structured (JSON + schema) representation. The structured data was optimized for the retrieval of Wiktionary senses and link types were categorized to produce concept definitions.

3.3 Concept Definitions

Formalization of the knowledge unit used in this work was done by firstly mapping each concept to a single Wiktionary sense. The concept is represented as a lexical/semantic graph, where a main addressing term, the root node, is connected to other terms through a set of edges. Each edge denotes a different type of lexical/semantic relationship, e.g. prefixation, synonymy/antonymy. The edges are also weighted, denoting the intensity of a relationship. Figure 3 shows a simplified visualization of a pair of different concept graphs for

[2] www.wiktionary.org

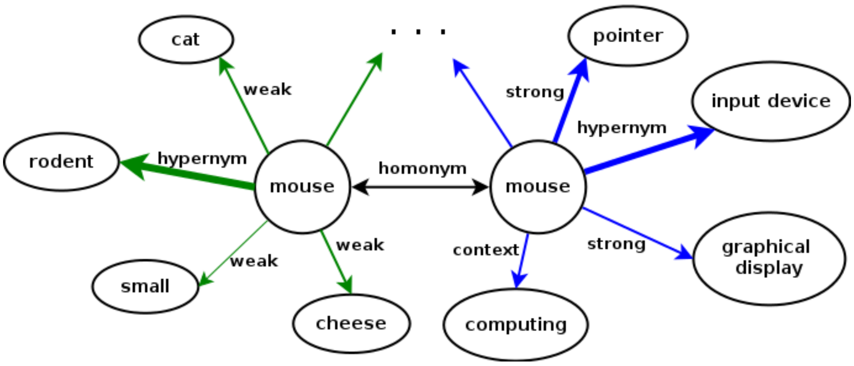
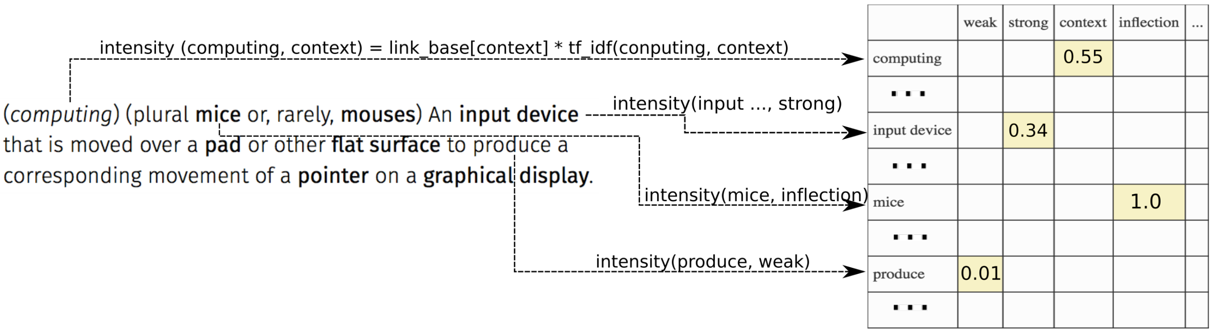


Figure 3: A simplified visualization of two concept graphs for the term “mouse”. The leftmost one denotes the concept of the small rodent and the other denotes the computer input device. The edge labels represent the relationship type and the thickness represent the its intensity.

Table 1: Link types used for the construction of concept graphs. They comprise both lexical (morphology, etymology) and semantic relationships between the root term, i.e., the Wiktionary entry title, and the terms used to describe the meaning.

Type	**Description**
weak	A term included in the description of the meaning on the Wiktionary entry.
strong	A term linked to another entry, i.e. a {highlight}, included in the description of the meaning.
context	A Wiktionary context link, explaining a specific situation in which the meaning described occurs.
synonym	A synonym relation. If it is an antonym, the sign of the link is reversed.
hypernym	A hypernym relation.
homonym	A homonym relation.
abbreviation	If the meaning described is given by interpreting the root term as an abbreviation.
etymology	Used to describe the origin of the root term of this meaning.
prefix	Denotes a prefixation (morphological) relationship of the root term.
suffix	Denotes a suffixation relationship. Same as above.
confix	Denotes a confixing relationship. Same as above.
affix	Denotes an affixation relationship. Same as above.
stem	Denotes a morphological stem relationship of the root term.
inflection	Denotes an inflectional relationship of the root term.


intensity (computing, context) = link_base[context] * tf_idf(conputing, context)
intensity(input ..., strong)
intensity(mice, inflection)
intensity(produce, weak)

(*computing*) (plural **mice** or, rarely, **mouses**) An **input device** that is moved over a **pad** or other **flat surface** to produce a corresponding movement of a **pointer** on a **graphical display**.

	weak	strong	context	inflection	...
computing			0.55		
...					
input device		0.34			
...					
mice				1.0	
...					
produce	0.01				
...					

Figure 4: Representation of one Wiktionary sense definition for ”mouse” as an encoded matrix: the *concept definition*. Each Wiktionary link is categorized and mapped to a vector space.

a single lexical entry of Wiktionary. The different link types are used to create a vector space, in which the edges of the definition graph are represented. Table 1 describes the link types used in this work.

Each concept graph is represented by a $M^{L \times T}$ matrix called *concept definition*, where L is the number of link types and T is the vocabulary size. The link intensities are defined for each type, by multiplying a manually defined constant $link\_base$ (a model parameter) by the TF-IDF score calculated for the vocabulary with respect to the type. Figure 4 illustrates the process.

Wiktionary entries also cover foreign terms, listing senses in the source language, e.g., English meanings of the French word "avec" in the English language section. Definitions for these terms are also included into the concept definition set. Additionally, translation links are provided for many sense definitions. Such links, as well as term redirections, i.e., distinct terms pointing to the same Wiktionary entry, are mapped to a single concept. This allows foreign terms to take advantage of the same concept graphs as the source language equivalents.

3.4 Definition Vectors

Finally, association between the concept definitions and terms is established by composition. This is done by simple element-wise sum and average of all concept definition matrices mapped to a Wiktionary entry. The resulting matrix is flattened in its row axis, i.e, rows are concatenated in order, producing a $L \times T$-dimensional sparse vector called *term definition vector*. If the term is not a Wiktionary entry, i.e., is out-of-vocabulary (OOV), a character n-gram-based attempt of morphological decomposition is done and if a complete morpheme match is found in the concept definition set, the matched concepts are composed for the OOV term. This decomposition attempt is done as follows:

1. For each character $c_i, i \in [0, n]$ in a OOV term of length n:
   - i Create empty list $morph\_cand$ of morpheme candidates.
   - ii Set index $j = i$.
   - iii Find Wiktionary entries that match c and add them to $morph\_cand$. For the first and last characters, include prefixes and suffixes in the search, respectively.
   - iv Concatenate c_{j+1} to c.
   - v Increment j.
   - vi Repeat from iii.

This will produce a sequence of n morpheme candidate lists. If a sequence produced by taking a single morpheme candidate from each list matches the entire OOV term, it is considered a candidate decomposition. If there are multiple candidate decompositions, the one with the shortest stem is selected.

Figure 5 illustrates an OOV composition for the nonexistent term "*unlockability*", which has a complete morpheme match in the concept definition set. If a complete morpheme match is not found, the term is considered a proper noun (if no POS is provided), and given a null (zero) vector.

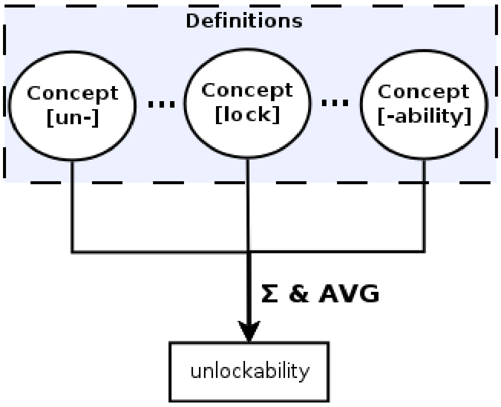


Figure 5: Morpheme match in concept definitions for the OOV term "*unlockability*".

Similarity comparison between two terms is done by measuring the cosine similarity between their definition vectors. A value closer to 1 indicates high similarity, a value closer to -1 indicates opposition and a value closer to 0 indicates unrelatedness. Table 2 shows a comparison table between semantic matches obtained using this method, *word2vec* (Mikolov et al., 2013) and *GloVe* (Pennington et al., 2014).

Table 2: Top closest and farthest to the term "*happy*" by Term Definition Vector, and closest Word2Vec (GoogleNews corpus), and GloVe (Wikipedia2014 + Gigaword) cosine similarities.

Def.Vec	**Def.Vec (-)**	**Word2Vec**	**GloVe**
joyous	sad	glad	glad
dexterous	unhappy	pleased	good
content	joyless	ecstatic	sure
felicitous	somber	overjoyed	proud
lucky	depressed	thrilled	excited

The definition vectors obtained in this way are also human interpretable to a certain extent. Each dimension corresponds to a link from the concept graphs used to compose a term definition. The values correspond to the strengths of such links. A human readable representation of the definition vector for the word "sunny", containing a maximum of two values per link type, can be written in the form: *weak@(a:0.0006, lot:0.003); strong@(cheerful:0.032, radiant:0.032); synonym@(bright:1.11, sunlit:1.11); suffix@(-y:1);*

stem@(sun:1); pos@(adjective:0.11, noun:0.11). In this example, $strong@radiant$ is a single vector dimension and the term is not disambiguated (multiple POS).

4 Experiments

4.1 Experimental setup

The definition vector representations obtained in this work were evaluated in the *SimLex-999* test collection for semantic similarity benchmark (Hill et al., 2015). This test collection contains a set of 999 English word pairs, associated to a similarity score given by a group of human annotators. The set is divided in 666 nouns pairs, 222 verb pairs and 111 adjective pairs. The Part-of-Speech (POS) information allows partial or complete disambiguation of the definition vectors. The choice of SimLex-999 was due to the type of similarity measured by this set, which excludes relatedness and is closer to the type of information captured by the concept definitions. Additionally, the *WordSim-353* (Finkelstein et al., 2001), *RG-65* (Rubenstein and Goodenough, 1965) and *MEN* (Bruni et al., 2014) test collections for semantic relatedness were also included in the evaluation, to verify the representation performance in measuring relatedness. While the MEN test collection also includes POS information, WordSim-353 and RG-65 do not include it, so sense distinction was not applied for the latter. Unfortunately, the test collections used in this work did not contain foreign words, so the translation-link features presented in Section 3.3 are solely presented as part of the method's description, and are not evaluated. This is due to the method being developed without focus on a specific test. The semantic similarity measurement was consequence of the method's design, but not its main target.

Evaluation is done by computing the Spearman's rank correlation coefficient (ρ) between the human annotators' similarity or relatedness scores and the scores given by the automated methods. A coefficient of value 1 means a perfect match between the relative positions of the pairs, when ranked by their similarity scores.

For the SimLex-999 test, the cosine similarity between the term definition vectors was set as the similarity score. For the WordSim-353, RG-65 and MEN tests, the absolute value of the cosine similarity was used instead, since opposite words are related. An additional test was performed to explore the possibility of combining distributional and definitional approaches. In this test, a small set of features was created to train a Learning-to-Rank model, in order to improve the similarity scores. The features were as follows:

- Presence of synonym, hypernym, strong and weak links [3] between the pair of words. Each link type is a separate feature.
- Term definition vector similarity.
- Word2vec similarity.

The features were computed for each pair and passed to SVM^{rank} (Joachims, 2006) for training and validation. A 10-fold cross-validation test using random pairs without replacement was run for the entire sets (5-fold for RG-65), except MEN. The MEN test collection is separated into training and testing sets, with 2K and 1K word pairs respectively, so these were used in place of the cross-validation. For each fold, the ranking scores provided by the trained ranker were used as similarity scores for calculating ρ. The average of all folds was considered the final result.

Experimental data and model parameters were set as follows:

- Linguistic information source: Wiktionary English database dump (XML + Wiki markup), 2015-12-26, containing more than 4 million entries. A reduced set, with only English, French, Greek, Japanese, Latin and Vietnamese language entries was used in the experiments. This set had about 734K entries, from which approx. 1 million concepts where extracted.
- $link\_base$ constants were set as: $weak = 0.2$, $strong = 2.0$, $context = 0.5$, $synonym = 10.0$, $hypernym = 5.0$, $homonym = 7.0$, $etymology = 1.0$ (also applied to morphological links) and $pos = 1.0$. The constants were adjusted by increasing or decreasing their values individually in intervals of 0.2, and observing the effect in ρ for SimLex-999 in the first fold of the cross-validation. The optimal values were selected and kept constant for the remaining folds and for the other tests. This was done because

[3] See Table 1

Table 3: Performance of different methods for the SimLex-999, WordSim-353, RG-65, and MEN test sets, reported as Spearman's rank correlation coefficient rho. The methods marked with $\diamond$ use a single information source. Fields marked with "-" indicate that the results were not available for assessment.

Method	ρ**@SimLex-999**	ρ**@WordSim-353**	ρ**@RG-65**	ρ**@MEN-1K**
Word2Vec (W2V) $\diamond$	0.38	**0.78**	0.84	0.73
GloVe $\diamond$	0.40	0.76	0.83	-
Term Def. Vectors (TDV) $\diamond$	0.56	0.36	0.68	0.42
Ling Dense	0.58	0.45	0.67	-
dLCE $\diamond$	0.59	-	-	-
TDV + W2V + SVM^{rank}	**0.62**	0.75	0.72	0.78
Recski et al. (2016)	**0.76**	-	-	-
ADW	-	0.75	**0.92**	-

changing $link\_base$ for each fold would create an unrealistic use scenario for our system, which cannot change $link\_base$ online. The cross-validation was repeated two times, with very minor differences between both test runs. The constant values reported here are from the last run.

- SVM^{rank} was set with a default linear kernel and C parameter (training error trade-off) was set to 8 for MEN and 5 for the other test collections. The value was increased in unit intervals, until convergence was longer than a time threshold (10 minutes). This parameter was adjusted using the training set for MEN, or inside each CV fold for the rest.

- Both *Word2Vec* and *GloVe* were used with pre-trained, 300-dimensional models: 100 billion words GoogleNews corpus and Common Crawl 42 billion token corpus respectively.

dLCE (Nguyen et al., 2016) was chosen as baseline, for being the best single information source method in the SimLex-999 test collection. Further results include Recski et al. (2016) (state-of-the-art), Ling Dense (Faruqui and Dyer, 2015), Word2Vec (Mikolov et al., 2013), and GloVe (Pennington et al., 2014). For WordSim-353, GloVe, Word2Vec, Ling Dense, and ADW (Pilehvar and Navigli, 2015) were included. For RG-65, Ling Dense, GloVe, and ADW (state-of-the-art), were included.

4.2 Results

The experimental results are presented in Table 3, where they are compared to other methods.

The results indicate that in the semantic similarity test, the term definition vectors perform closely to other representation models taking advantage of curated data, such as WordNet. It also outperforms the most popular distributional representations. However, they are clearly outclassed in the semantic relatedness test, for which the distributional approaches show superior performance.

An interesting observation can be made when combining word2vec similarity with term definition features through the use of Machine Learning. A performance trade-off seems to exist at the semantic relatedness tests, but the same is not true for the similarity test. This allowed the combined model to improve considerably at little cost. Further analysis helped in understanding the reason for this particularity (Section 4.3).

Lastly, the experiments have also shown that the method for extracting concept definitions is not computationally expensive. The developed implementation took about 6 minutes to extract all concept definitions from the structured Wiktionary data used in the tests, using a modern desktop computer (3GHz processor and at least 8GB RAM). Structuring Wiktionary data took less than 20 minutes with the same equipment, and was done a single time.

4.3 Error analysis

Identifying the flaws in a method is a fundamental step in improving it and also in understanding the problem it tries to solve. With this in mind, the error cases identified in measuring similarity from the SimLex-999 set were observed in detail. In this analysis we considered as error any word pair that was put among the top 15% similarity scores by the human annotators, but was ranked in the lower 50% using the definition vectors. The same applies for the bottom 15% scored by humans, that are ranked in the upper half by our approach.

The errors found were classified in four categories:

- Insufficient links in Wiktionary: this type of error occurs when the wiktionary sense corresponding to a concept lacks annotations. Typical cases contain only a short description, with no links. The concept graph is then left with only weak links, which have little impact on similarity calculation. The pair *drizzle–rain* (noun) is one example of this.
- Undeclared hypernymy: certain cases of hypernymy are not solved in the concept extraction, since they require multiple hops in the definition links to be found. The pairs *cop–sheriff* and *alcohol–gin* (noun) are instances of such problem.
- Casual vs. formal language semantics: not a flaw in the method per se, but an error caused by the differences in formal description of a language (in a dictionary), when compared to casual use. The pair *noticeable–obvious* (adjective) illustrates this.
- Other: flaws in the extraction process or annotation problems in Wiktionary.

Those errors affect the pairs in the top 15% human similarity scores 7 times more than the lower 15%. They are distributed as shown in Table 4.

Table 4: Distribution of definition vector error types in SimLex-999.

Type of error	Proportion
Insufficient links	21.4%
Undeclared hypernymy	38.1%
Casual semantics	14.3%
Other	26.2%

Having about one quarter of the errors in the “other” category shows that there is some space for improvement in the concept extraction process. The insufficient links and undeclared hypernymy categories are cases in which distributional approaches may do better if similarity is high, due to the words intrinsic relatedness.

Analysis of SVM^{rank} scores showed that the insufficient links category benefited the most from the combination with word2vec. The reason is that the features chosen for use with the ranker made such cases distinguishable and more likely to receive a larger weight from the word2vec similarity score after training. The undeclared hypernymy cases, on the other hand, are not so evident and would require a more complex approach on the concept extraction process.

5 Conclusion

Alternative approaches to distributional language representations can take advantage of information unavailable through pure statistical means. Taking advantage of large curated linguistic resources is a popular way of obtaining such information and offers large room for exploration. With this in mind, we propose a novel method for obtaining vector representations of lexical items using Wiktionary sense definitions. The lexical item representations are composed from basic meaning units called concept definitions, which are extracted from the linguistic resource. Results obtained from a semantic similarity evaluation test indicate performance comparable to other methods based on linguistic resource extraction. Furthermore, a noticeable performance gain was obtained by applying a Machine Learning approach to combine word2vec similarity scores with the ones obtained using this approach, exceeding both methods’ results.

Planned improvements include the use of a graph traversal approach to capture deeper semantic links and also the inclusion of translation links as separate dimensions in the concept vector space, in order to facilitate the use of obtained representations as a machine translation resource. The inclusion of distributional similarity measures as separate dimensions is also under study and provides an alternative way of combining the distributional and definitional approaches.

Acknowledgments

We also would like to thank anonymous reviewers for their detailed comments for improving our paper. This work was partially supported by CNPq (National Counsel of Technological and Scientific Development) - Brazil, JSPS KAKENHI Grant number 15K16048, JSPS KAKENHI Grant Number JP15K12094, and CREST, JST.

References

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 809–815. Association for Computational Linguistics.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research (JAIR)*, 49(1-47).

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Joaquín Derrac and Steven Schockaert. 2015. Inducing semantic relations from conceptual spaces: a data-driven approach to plausible reasoning. *Artificial Intelligence*, 228:66–94.

Cícero Nogueira dos Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *ICML*, pages 1818–1826.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

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

Christiane Fellbaum et al. 1998. Wordnet: An electronic database.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.

Taher H. Haveliwala. 2002. Topic-sensitive pagerank. In *Proceedings of the 11th international conference on World Wide Web*, pages 517–526. ACM.

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

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

Thorsten Joachims. 2006. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 217–226. ACM.

András Kornai. 2010. The algebra of lexical semantics. In *The Mathematics of Language*, pages 174–199. Springer.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*.

Anh Kim Nguyen, Sabine Schulte im Walde, and Thang Ngoc Vu. 2016. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 454–459. Association for Computational Linguistics.

Charles Kay Ogden, Ivor Armstrong Richards, Sv Ranulf, and E Cassirer. 1923. The meaning of meaning. a study of the influence of language upon thought and of the science of symbolism.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

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

Gábor Recski, Eszter Iklódi, Katalin Pajkossy, and Andras Kornai, 2016. *Proceedings of the 1st Workshop on Representation Learning for NLP*, chapter Measuring Semantic Similarity of Words Using Concept Networks, pages 193–200. Association for Computational Linguistics.

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

Magnus Sahlgren. 2008. The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):33–54.

Miran Seok, Hye-Jeong Song, Chan-Young Park, Jong-Dae Kim, and Yu-seop Kim. 2016. Named entity recognition using word embedding as a feature. *International Journal of Software Engineering and Its Applications*, 10(2):93–104.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008a. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *LREC*, volume 8, pages 1646–1652.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008b. Using wiktionary for computing semantic relatedness. In *AAAI*, volume 8, pages 861–866.