2017 9th International Conference on Knowledge and Systems Engineering(KSE)

WikTDV: Data extraction and vector representation resource for Wiktionary senses

Danilo S. Carvalho
School of Information Science
Japan Advanced Institute of Science and Technology
Nomi City, Japan 923-1292
Email: danilo@jaist.ac.jp

Minh-Le Nguyen
School of Information Science
Japan Advanced Institute of Science and Technology
Nomi City, Japan 923-1292
Email: nguyenml@jaist.ac.jp

***Abstract*—Effective use of collaborative web resources, such as Wikipedia and Wiktionary, has been a recurrent topic of research in the Natural Language Processing and Information Retrieval communities. The same can be said about the use of vector-based language representations, e.g., word, sentence, document embeddings. However, there is currently a shortage of resources that offer vector representations that can take advantage of the structural properties of web resources. This paper describes a system for extracting information from Wiktionary to a machine-readable format and using this information to obtain vector representations that can be used for semantic similarity computation and basic word sense disambiguation. The methodology used to build the system is also discussed. Experimental evaluation on the semantic similarity task indicate efficiency close to the reference method applied in this work. A web service and visualization facilities complete the set of contributions.**

I. Introduction

Collaboratively built World Wide Web resources, such as *Wiktionary* [1] have become a prominent source of information for Natural Language Processing (NLP) and Information Retrieval (IR) research in recent years. Reasons for their use include the massive scale of information coverage on multiple domains and languages, with a steady growth rate [2], as well as their ready availability and open data access policies. Such resources, however, are directed towards human interaction (for reading and editing) and are not structured for computer use. Therefore, several research efforts were undertaken to obtain computer-readable, i.e., *structured*, data from them, extract relevant information and leverage it for a variety of NLP tasks. Gabrilovich and Markovitch [1] proposed a method to represent textual meaning as a weighted vector of *Wikipedia* [3] based concepts, having each Wikipedia article to be a single dimension in the resulting vector space. Zesch et al. [2] presented application programming interfaces (APIs) for structured access of Wiktionary data, and Krizhanovsky and Smirnov [3] proposed an approach for building a general-purpose lexical ontology from the same data. Liebeck and Conrad [4] presented a parser for Wiktionary that collects morphological information, which is used for improving lemmatization tasks. Aouicha et al. [5] proposed a combined taxonomic measurement method for calculating *semantic relatedness and similarity*, using *WordNet* [6], Wikipedia categories and Wiktionary data. Carvalho and Nguyen [7] proposed a method for obtaining vector representations from Wiktionary senses, based on lexical and link data weighting. To minimize information quality issues, expert curated knowledge bases, such as *WordNet* [6], are often used in conjunction with the web resources.

Text-based Information Retrieval (IR) and Knowledge Engineering (KE) research has been employing collaborative web resources information to build and exploit knowledge graphs [8] with success. Concurrently, interest in vector-based language representations is growing recently [9], [10]. Techniques based on the *Distributional Semantic Hypothesis* such as *ESA* [1], *Word2Vec* [11] and *GloVe* [12] allow the inclusion of contextual semantic information from vast ammounts of unstructured textual data into a variety of IR & KE approaches, in particular Machine Learning based ones. However, while several resources are available for vector representations based on unstructured sources (text embeddings), there is a lack of resources for vector representations built upon structured data. Such representations could help NLP, IR and KE researchers to combine benefits of collaborative web resources and expert curated information with distributional and distributed semantics techniques.

This paper describes a combined parsing, information extraction and structuring system for *Wiktionary* data, as well as the methodology employed in its construction. 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. The extracted information is in the form of preprocessed semi-structured data and vector-based language representations of Wiktionary entries. Semantic similarity computation and word sense disambiguation functionalities were also developed, and are processed using the extracted information. Our contributions are:

1) An event-driven parser for the Wiktionary database file format that generates semi-structured (JSON [4]) data.

[1] www.wiktionary.org
[2] https://stats.wikimedia.org/wiktionary/EN/ChartsWikipediaEN.htm
[3] www.wikipedia.org
[4] JavaScript Object Notation

978-1-5386-3576-6/17/$31.00

2) An information extraction system for obtaining multi-language, vector-based representations from Wiktionary sense data.
3) A RESTful web service for processing user inputs and serving resources derived from (1) and (2).
4) A visualization interface for (3).

The system developed in this work and data obtained from it are freely available to the public as open source code and open access data [5].

The remainder of this paper is organized as follows: Section II presents related work and comparisons with the proposed resource. Section III explains the parsing approach adopted for Wiktionary, followed by an explanation of the information extraction approach in Section IV. Section V describes the resources available in the proposed system and shows an experimental evaluation on the semantic similarity task. Finally, Section VI offers some concluding remarks.

II. Related work

Efforts for a public available extraction mechanism of structured data from Wiktionary can be found in the works of Zesch et. al [2], and Krizhanovsky and Smirnov [3]. The former (JWKTL) is distributed as a code library (Java) and the latter (Wikokit) is distributed as a standalone application and pre-parsed database files. In common, they generate relational databases with the extracted data, which is then exposed through an application programming interface (API). As a consequence, they both require a DBMS[6] to be used and the database tables need to be updated for any change in the underlying extraction schema, e.g. adding new fields for an Wiktionary entry. A similar mechanism is offered for WordNet by the NLTK [7] library, but without using a relational database. The APIs offer a simplified method for accessing page, entry and sense data as text literal lists and link sets.

A recently developed method for creating vector-based language representations from Wiktionary data can be found in the work of Carvalho and Nguyen [7]. In [7], links are categorized into different types and used to produce a weight matrix for each sense entry. Such matrices are then manipulated to generate sparse vectors for either senses or combinations thereof, representing ambiguity. Those representations are called *term definition vectors* and their approach is named as *definitional* instead of *distributional*, since it does not employ distribution over usage context.

The system developed in this work does not generate a relational database. Instead, the Wiktionary parser produces a semi-structured, document model file (Section III), that can be read by any JSON-campatible programming environment or library. This file is also human-readable, and can be searched by text stream utilities, such as *GNU Grep* [8]. Additionally, the information extraction system generates vectors as proposed in [7], but also including link types for translations, redirections and back-translations (non-specific sense links from translated terms back to the source language). With the obtained vectors, the system is able to calculate semantic similarity, with the possibility of "disambiguating" the vectors by including additional terms as context or part-of-speech (POS) tags.

III. Wiktionary database parsing

The data available from Wiktionary is composed of a set of markup documents, one for each entry. All the documets are included in a single "database dump file" (XML format). The extraction procedure is divided in two stages: parsing and information extraction. Parsing takes as input an English Wiktionary database dump file and is done in three steps:

1) Count the number of entries and register their file offset.
2) Create a specified number of parsing processes and assign a group of entries for each process.
3) Collect and merge the results of each parsing process to generate the output JSON file.

Each parsing process takes as input Wiktionary entries, which are single XML elements, containing metadata and the contents of the entry (part-of-speech and sense data, examples, etymology, among others) in *Wiki markup* [9] format. The entry contents are parsed line-by-line, in event-driven fashion. Each line is checked for identifying patterns through regular expression matching, and upon finding a relevant pattern, the corresponding routine for extracting formatted data is called. For each entry, a JSON document is created with the structure illustrated in Figure 1.

After all entries have been processed, the parser results are merged into a single document list and sorted by the entry title. This document list is the parsing final output. An example output of this stage was made available for download [10], containing over 730K entries.

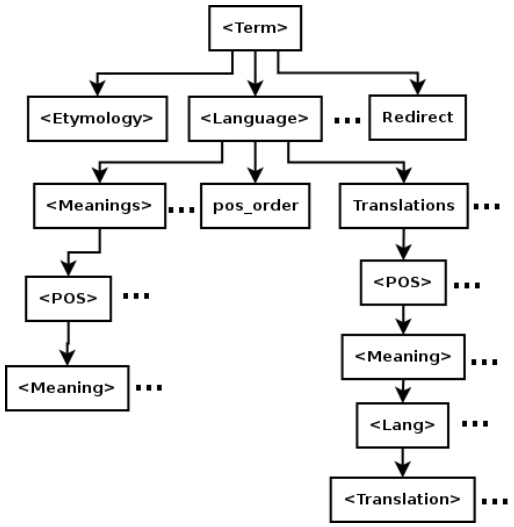

Fig. 1. Wiktionary entry hierarchical structure obtained from the parser.

[5]https://github.com/dscarvalho/tdv
[6]Database Management System
[7]http://www.nltk.org/howto/wordnet.html
[8]https://www.gnu.org/software/grep/
[9]https://en.wiktionary.org/wiki/Help:Wikitext_quick_reference
[10]https://goo.gl/bkVhoq

IV. Information Extraction - Term Definition Vectors

The information extraction stage takes as input the JSON output from the parser and is done in a single step. It uses the description, link markup, morphological and semantic information from each sense in each entry to produce a typed link weight matrix (Figure. 2) called *concept matrix*, as described in [7]. Table I describes the link types used. The matrix is then flattened by concatenation of its rows (vocabulary dimension) to obtain *concept vectors*. For a given term, its concept vectors can be combined (summed) to represent ambiguity. A combination of concept vectors is called *term definition vector*. The concept vectors are cached in memory or disk and are indexed by their entry title and POS or by a hash based sense index, so they can be quickly retrieved. Figure 4 illustrates the entire information extraction process flow.

	weak	strong	...	hypernym	...
furniture		0.34		0.8	
...					
legs	0.06				
...					
sofa		0.25			

Fig. 2. Link weight matrix for the first Wiktionary sense of the term "chair" (noun). The total dimension of the matrix is $V \times T$ where V is the vocabulary size and T is the number of link types.

	furniture				...	legs			
	weak	strong	...	hypernym	...	weak	strong	...	hypernym
...		0.34		0.8		0.06			

Fig. 3. After flattening, the concept matrix is turned into a $1 \times (T * V)$ sparse vector, called concept vector.

V. Available Features

To facilitate the use of the extracted information, a web service was developed, exposing the following functionality:

A. Structured Wiktionary entry

Given a single term (a word or short phrase), provides the semi-structured Wiktionary data obtained from the parser, added of the meaning indexes. The following example shows

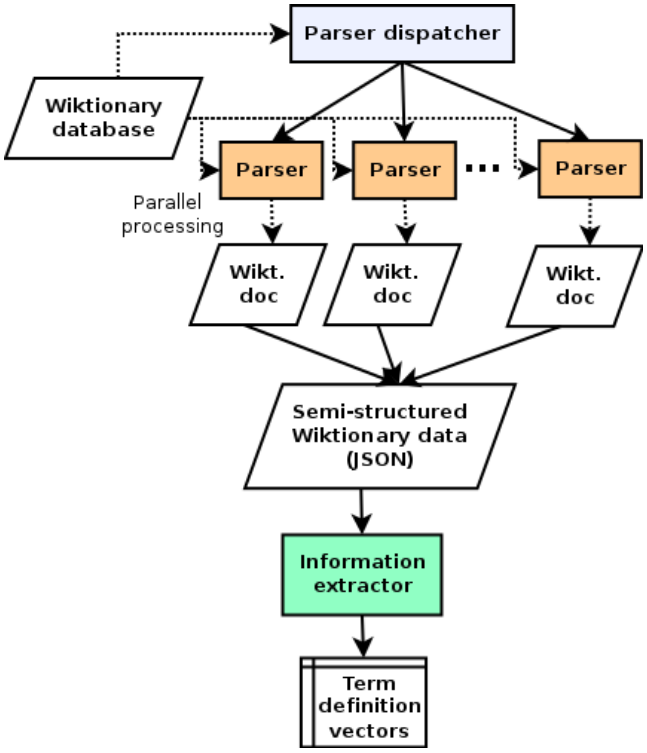

Fig. 4. Flow of the information extraction process.

a fragment of the output data for the term "today". A full example was made available for download [11].

```
{ "langs" : {
      "English" : {
         "pos_order" : [
            "adverb",
            "noun"
         ],
         ...

         "meanings" : {
            "noun" : [
               { "links" : [
                     "current",
                     "day",
                     "date"
                  ],
                  "id" : 6171340000003,
                  "meaning" : "A current day or date"
               }
            ],
            ...
}
```

B. Term Definition Vector representation

Given a term, an optional part-of-speech (POS) and context (a word list), provides the term vector definition representation [7] as a set of typed links between the provided term and other Wiktionary entries. When POS is given, the senses are filtered by the provided POS, given that each Wiktionary sense has POS and this information is indexed together with

[11] https://goo.gl/QS1hzs

TABLE I
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.
translation	Denotes a translation link relationship (target language) to the root term (source language).

the definition vectors. The POS filtering generates a less ambiguous vector, with fewer links. When a context is given, vectors are obtained for each context word in the same fashion, and are then used as a cluster, from which the average semantic similarity is calculated (Section V-C) for each sense of the term to be vectorized. The vector with the lowest average distance is chosen as the output. Figure 5 illustrates the vector selection with context.

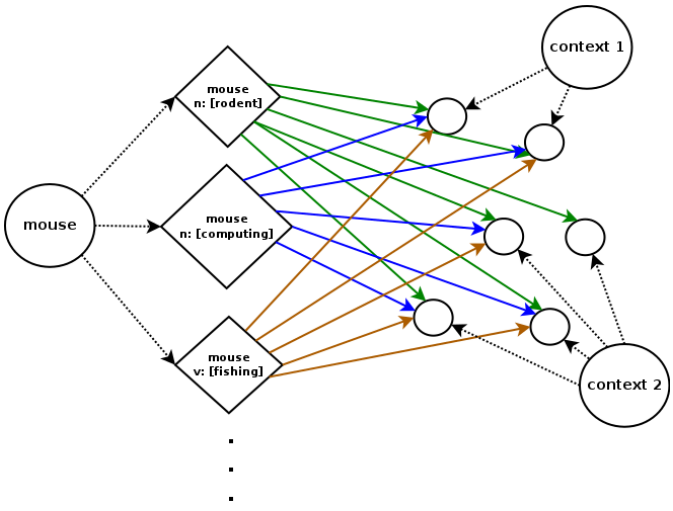

Fig. 5. Term disambiguation by context cluster similarity for two distinct context, given the term "mouse". The distances (cosine similarity) of the different senses: "rodent", "computing", "fishing" are calculated with respect to a set of terms of each context (the blank circles). The concept with lowest average distance is selected.

The vectors can be also provided in a "machine-friendly" format, which contains only the link indexes and weights. This format can be used to obtain sparse vector features compatible with the *SVM-Light* [12] Support Vector Machine implementation. Example outputs for both formats were also made available for download [13] [14].

C. Semantic similarity

A semantic similarity measure is obtained through simple computation of cosine. Tests were conducted in the *SimLex-999* [13] test collection for semantic similarity benchmark and the *MEN* [14] test collection for semantic relatedness.

The SimLex-999 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 MEN test collection consists of 3000 word pairs and their human-assigned similarity score. The pairs are randomly selected from words that occur at least 700 times in the ukWaC and Wackypedia corpora combined [15] and at least 50 times (as tags) in the opensourced subset of the ESP game dataset [16]. Sampling is done so that the pairs represent a balanced range of relatedness levels according to a text-based semantic score.

Tests were run with *link_base* constants: $weak = 0.2$, $context = 0.5$, $etymology = prefix = suffix = confix = affix = stem = inflection = 1.0$ and $pos = 1.0$, $strong = 2.0$, $hypernym = 5.0$, $homonym = 7.0$, $translation = 7.0$, $synonym = abbreviation = 10.0$. A comparison was made with *Word2Vec* [11] and *TDV* results reported in [7]. The results are presented in Table II. The decrease in performance when compared to the reference method [7] is due to small differences in the generated concept vectors, which in this work include translation links. This difference allows a multi-language vector space at a small expense of single language performance.

[12]http://svmlight.joachims.org/

[13]https://goo.gl/ldEX4d
[14]https://goo.gl/L5ZuB1
[15]http://wacky.sslmit.unibo.it/
[16]http://www.cs.cmu.edu/~biglou/resources/

TABLE II
SPEARMAN'S RANK CORRELATION COEFFICIENT *rho* FOR THE SEMANTIC SIMILARITY RANKING ON THE SIMLEX-999 AND MEN TEST SETS.

Method	ρ**@SimLex-999**	ρ**@MEN-1K**
Word2Vec [11]	0.38	0.73
TDV [7]	0.56	0.42
Our resource	0.54	0.39

D. Word sense disambiguation

A best-effort attempt at word sense disambiguation is also provided, using the same mechanism of vector disambiguation described in Section IV. It takes as input a sentence and a chosen term from the sentence, with an optional POS. Figure 7 in Section V-E shows an usage example of this feature. All terms of the sentence, except the selected one, are used as context for the disambiguation.

E. Visualization interface

Aiming to present the resource in a form easier to understand and manipulate, a simple visualization interface was developed for the vector representation and word sense disambiguation functionalities. Figures 6 and 7 show the appearance of the interface. The purpose of the graph representation is to give an loose idea of the vector composition of a term. A precise and readable vector composition can be obtained by not using the "machine-friendly" format for the output vectors. The word sense disambiguation interface facilitates the use of the corresponding web service functionality, by providing POS disambiguation choices and the choice of term to disambiguate through automatic sentence tokenization.

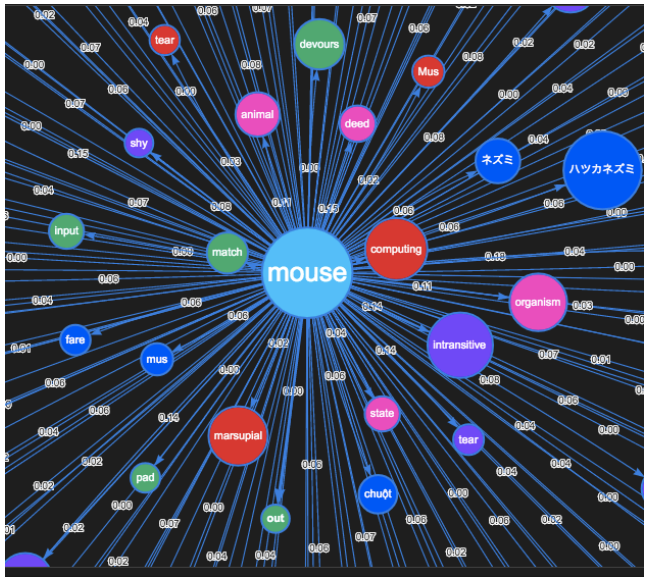

Fig. 6. Graph representation of the term definition vector. Each edge represent a single (non-zero) dimension of the vector space, which is composed by the different types of links to Wiktionary entries. The distance to the center (root) node is a relative approximation of the weights, with respect to the other nodes. The size of the nodes and angle between edges have no special meaning.

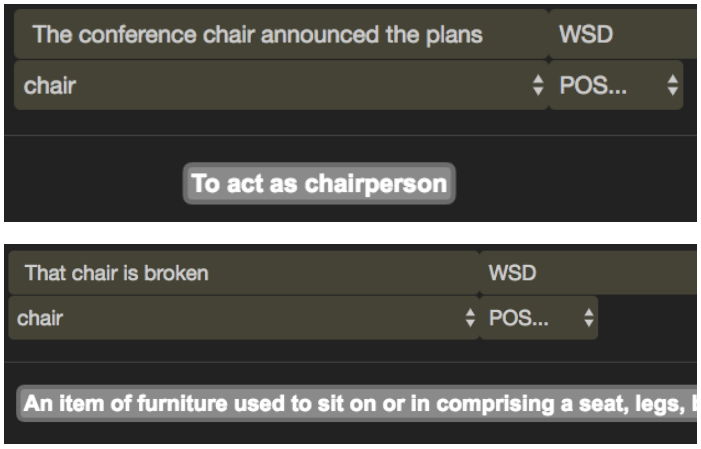

Fig. 7. Example of word sense disambiguation. The correct meaning of the word "chair" is identified for both sentences.

VI. CONCLUSION

While collaboratively built language resources and vector-based language representations have seen increased interest by the NLP and IR research communities, there is a lack of resources combining strengths of both resource types. This work described a system that provides semi-structured data for the collaborative web dictionary + thesaurus Wiktionary, and vector representations built up upon this data. Experimental evaluation on semantic similarity indicates consistency regarding the implementation's reference method, altough with some space for improvement.

Immediate future concerns for this work are the preparation of infrastructure for providing the resources and improvement of both semantic similarity computation and the visualization interface. Building links to a open Wiktionary ontology would further improve the usefulness of this resouce for IR & KE research and is also planned.

ACKNOWLEDGMENTS

This work was partially supported by JSPS KAKENHI grant number 15k16048 and CNPq (National Counsel of Technological and Scientific Development) - Brazil.

REFERENCES

[1] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis." in *IJcAI*, vol. 7, 2007, pp. 1606–1611.

[2] T. Zesch, C. Müller, and I. Gurevych, "Extracting lexical semantic knowledge from wikipedia and wiktionary." in *LREC*, vol. 8, no. 2008, 2008, pp. 1646–1652.

[3] A. A. Krizhanovsky and A. V. Smirnov, "An approach to automated construction of a general-purpose lexical ontology based on wiktionary," *Journal of Computer and Systems Sciences International*, vol. 52, no. 2, pp. 215–225, 2013.

[4] M. Liebeck and S. Conrad, "Iwnlp: Inverse wiktionary for natural language processing," in *ACL (2)*, 2015, pp. 414–418.

[5] M. B. Aouicha, M. A. H. Taieb, and A. B. Hamadou, "Taxonomy-based information content and wordnet-wiktionary-wikipedia glosses for semantic relatedness," *Applied Intelligence*, vol. 45, no. 2, pp. 475–511, 2016.

[6] G. Miller and C. Fellbaum, "Wordnet: An electronic lexical database," 1998.

[7] D. S. Carvalho and M. L. Nguyen, "Building lexical vector representations from concept definitions," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017) (preprint)*. Association for Computational Linguistics, 2017. [Online]. Available: http://www.jaist.ac.jp/~s1520009/preprint/eacl2017_carvalho_nguyen.pdf

[8] K. Balog and R. Neumayer, "A test collection for entity search in dbpedia," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2013, pp. 737–740.

[9] D. Ganguly, D. Roy, M. Mitra, and G. J. Jones, "Word embedding based generalized language model for information retrieval," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2015, pp. 795–798.

[10] F. Diaz, B. Mitra, and N. Craswell, "Query expansion with locally-trained word embeddings," in *Proceedings of Neu-IR: The SIGIR 2016 Workshop on Neural Information Retrieval (preprint)*, 2016. [Online]. Available: https://arxiv.org/abs/1605.07891

[11] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, 2013.

[12] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014, pp. 1532–1543. [Online]. Available: http://aclweb.org/anthology/D14-1162

[13] F. Hill, R. Reichart, and A. Korhonen, "Simlex-999: Evaluating semantic models with (genuine) similarity estimation," *Computational Linguistics*, vol. 41, no. 4, pp. 665–695, 2015. [Online]. Available: http://aclweb.org/anthology/J15-4004

[14] E. Bruni, N.-K. Tran, and M. Baroni, "Multimodal distributional semantics," *Journal of Artificial Intelligence Research (JAIR)*, vol. 49, no. 1-47, 2014.