EVALution-MAN: A Chinese Dataset for the Training and Evaluation of DSMs

Hongchao Liu, Karl Neergaard, Enrico Santus, Chu-Ren Huang
CBS, The Hong Kong Polytechnic University
11 Yuk Choi Rd. Hunghom, Hong Kong
jiye12yuran@126.com, {karlneergaard, esantus}@gmail.com, churen.huang@polyu.edu.hk

Abstract

Distributional semantic models (DSMs) are currently being used in the measurement of word relatedness and word similarity. One shortcoming of DSMs is that they do not provide a principled way to discriminate different semantic relations. Several approaches have been adopted that rely on annotated data either in the training of the model or later in its evaluation. In this paper, we introduce a dataset for training and evaluating DSMs on semantic relations discrimination between words, in Mandarin, Chinese. The construction of the dataset followed EVALution 1.0, which is an English dataset for the training and evaluating of DSMs. The dataset contains 360 relation pairs, distributed in five different semantic relations, including antonymy, synonymy, hypernymy, meronymy and nearsynonymy. All relation pairs were checked manually to estimate their quality. In the 360 word relation pairs, there are 373 relata. They were all extracted and subsequently manually tagged according to their semantic type. The relatas'frequency was calculated in a combined corpus of Sinica and Chinese Gigaword. To the best of our knowledge, EVALution-MAN is the first of its kind for Mandarin, Chinese.

Keywords: Dataset, Distributioanl Semantic Models, Tarining, Evaluating

1. Introduction

Distributional semantic models (DSMs) have been applied to both the measurement of semantic similarity and relatedness, and are of paramount importance for tasks such as word sense disambiguation, lexical replacement, dictionary construction, and entailment understanding, to name a few (Sun, 2014). DSMs are founded on the assumption that the lexical similarity between words depends on their distributed context. According to the *Distributional Hypothesis*, if compared concepts or words have similar distributional features (hence, context), they are likely to have higher semantic similarity (Harris, 1954). One major shortcoming of current DSMs is that they cannot be used for discrimination among different relationships between words (Santus et al., 2014c).

In recent years, DSMs have been adopted in several semantic tasks, including the identification of semantic relatedness and similarity. The former task is concerned with whether two words are related or not, independently from their paradigmatic similarity. The identification of semantic similarity, instead, is a more specific task and it consists in identifying words that are paradigmatically related (Sun, 2014; Murphy, 2003). DSMs have successfully addressed these two tasks by using vector cosine as a measure of similarity (Agirre et al., 2009).

Unfortunately, however, DSMs are not yet able to discriminate the several semantic relations that exist between words (Santus et al., 2014c). For example, a word such as 龍 (*long*, dragon) can be in relation with several other words:

1. Antonymy: 龍 (*long*, dragon) vs. 鳳 [1] (*feng*, phoenix)
2. Hypernymy: 龍 (*long*, dragon) vs. 水中生物 (*shuizhongshengwu*, aquatic life)
3. Nearsynonymy [2]: 龍 (*long*, dragon) vs 兔 (*tu*, rabbit)
4. Synonymy: 不能 (*buneng*, cannot) vs. 不可以 (*bukeyi*, cannot/not able to)
5. Meronymy: 墻 (*jianzhuwu*, wall) vs. 建築物 (*jianzhuwu*, building)

A DSM might be able to identify these words as similar/related, but it will struggle to to discriminate the paradigmatic relations they hold. While this is indeed still a challenging task, the NLP community has adopted several approaches (Jia et al., 2014). The new approaches can be classified according to the need of annotated resources to achieve their goal of identifying multiple relations between given word pairs: *supervised*, which is a method of discrimination that relies on large scale datasets to train the models (Girju et al., 2006; Van Hage et al., 2006); *semi-supervised*, which uses a small dataset that acts as a seed to extract more relation instances and patterns iteratively (Pantel and Pennacchiotti, 2006); and *unsupervised* which does not require any manually annotated dataset to train the model (Jia et al., 2014; Santus et al., 2014a; Santus et al., 2014b; Santus et al., 2014c).

While a dataset is necessary for training the former two approaches, the third approach will also need a dataset for testing. Because of this necessity, several banchmarks have been constructued for English. To meet the demand, several benchmarks have been constructed for English *TOEFL* (`Landauer and Dumais, 1997`), *BLESS* (`Baroni and Lenci, 2011`), LENCI/*Benotto* (`Benotto, 2015`) and *EVALution 1.0* (`Santus et al., 2015`). However, to date there is no dataset especially designed for DSMs in Mandarin, Chinese.

In this paper, we describe *EVALution-MAN*, a new resource for training and evaluating Mandarin DSMs. EVALution-MAN is a traditional Chinese version of the English *EVALution 1.0* (Santus et al., 2015), as it was built following a very similar methodology.

[1] In Chinese culture, dragon and phoenix roughly mean male and female, or king and queen.

[2] This term, defined in Chinese WordNet, implies relatedness instead of similarity.

2. Related Work

2.1. Datasets for DSMs

For the training and evaluation of DSMs, several datasets have been widely used. While general-purpose datasets are still being made use of, a recent trend in the creation of benchmarks for the training of DSMs has been their construction from data derived from specific tasks performed by human participants (Hill et al., 2015).

The general purpose resources used for the training and evaluation of DSMs have followed the model laid out by WordNet (Fellbaum, 1998). For Mandarin, Chinese, there is Chinese WordNet (CWN: (Huang et al., 2010)) and How-Net (Liu and Li, 2002). CWN consists of more than 50,000 word relation pairs covering the relationships of antonymy, synonymy, hyponymy/hypernymy, meronymy/holonymy, paranymy, nearsynonymy and variant. Paranymy in this context refers to co-hyponymy, i.e., lexical items that share hypernym pairs, while variant refers to pairs that are identical in meaning and use but differ in orthography (Huang et al., 2010). Nearsynonymy here refers to two words are related instead of similar. Because CWN was constructed using WordNet, it carries all its limitations (e.g. arbitrariness). This recently has been criticized, as it does not carry any information about human judgments (Santus et al., 2015).

How-Net (Liu and Li, 2002) is constructed as an interconnected graph of relations that span synonymy, antonymy, hypernymy/hyponymy, and semantic argument information such as agent, event, patient, location, etc. While structurally distinct from WordNet, the relation pairs that can be extracted from its graph structure also have not been evaluated against human judgment.

A well-known benchmark for the evaluation of DSM is the set of eighty multiple choice synonym questions from the Test of English as a Foreign Language (TOEFL). This dataset was introduced for the first time by (Landauer and Dumais, 1997) and it allowed for the comparision of computer performance (and other computational models) against that of the performance of college applicants. (Mohammad et al., 2008) used a similar paradigm for their dataset, built from 162 questions from the Graduate Record Examination (GRE) that targeted antonymy. Both datasets address only one semantic relation. Their sizes and focus on single relations make them inappropriate for an extensive evaluation of DSMs.

BLESS (**B**aroni and **L**enci **E**valuation of **S**emantic **S**paces) was the first English dataset especially designed for the evaluation of multiple semantic word relations. It features 200 basic target concepts instantiated by 26,554 relata. It contains five different word relations: co-hypernymy, hypernymy, meronymy, attribute, and event. The structure of the dataset is in the form of "concept-relation-word"tuples. For each concept, BLESS also features semantically unrelated words. The shortcoming of *BLESS*, however, is that the dataset didn't take synonymy and antonymy into consideration.

Benotto (2015) constructed an English dataset targeting hypernymy, synonymy, and antonymy through elicitation experiments following the method introduced by (Paradis et al., 2009). Target words were firstly selected from sources such as WordNet and GermaNet (Hamp et al., 1997). An elicitation experiment was then conducted through Amazon Mechanical Turk to ask every participant to produce antonyms, synonyms and hyponyms of each word. In this way, 8,910 word relation pairs were collected.

A DSMs-oriented resource that overcomes the shortcomings of the above datasets for English is EVALution 1.0 (Santus et al., 2015). It consists of 7,429 word relation tuples including 1,829 relata. It contains seven relation types including hyponymy, antonymy, synonymy, meronymy, entailment (if X is true, Y is true), possession (has a) and attribute. Each word relation pair was validated manually in a sentence judgment task by 5 participants using crowdsourcing website Crowdflower. The relata were then tagged in a second task that asked the subjects to tag each words'semantic context/domain; for example, whether the word could be described as an event, space, an object, emotion, food, etc.

2.2. Mandarin Chinese Word Relation Studies

Two approaches have been adopted for the measurement of word similarity in Mandarin, Chinese: knowledge-based methods and those implementing corpora.

Wang (1999) measured word similarity according to the distance between word pairs constructed according to a thesaurus (Mei et al., 1983), wherein distance refers to the number of nodes (words within the tree structure) that must be traversed along the path between each target word. The difficulty with such a method is whether there is any reliability to the conceit that the less nodes the path traverses, the more similar two words are (Sun, 2014). Meanwhile, Liu and Li (2002) used How-Net as the dataset for their measurement of word similarity. As mentioned previously, How-Net is not constructed in the same way as CWN or WordNet. All of the words in How-Net are described by the 1,500 basic sememes (semantic features). There are no direct relations between words, and relations only exist between basic semantic elements. This makes the calculation of word similarity as done by Wang (1999) not directly translatable. Thus Liu and Li (2002) calculated word similarity based on sememe similarity. There are two possible problems with this study. Firstly, the study only considered a single relation: hyponymy. Secondly, relations between sememes might not be equivalent to relations between words.

In a final study that implemented unsupervised methods Jia et al. (2014) tried to acquire the part-whole relation (i.e., meronymy). The paper especially pointed out that there is currently no dataset detailing part-whole relations that can be used for the evaluation of their methodology for Mandarin. The current paper addresses the necessity and importance of the construction of a practical dataset for the training and evaluation of DSMs.

3. Construction of Dataset

The word pairs from which *EVALution-MAN* was constructed came from Chinese WordNet (CWN: (Huang et al., 2010)). CWN is a knowledge system modeled on the original Princeton WordNet. The system's word

sense examples and lexical semantic relations came from Sinica Corpus (Chen et al., 1996). We extracted pairs holding the following relations: synonymy (i.e. words belonging to the same CWN synset); antonymy (i.e. words that have the opposite synset in CWN); hyponymy (i.e. one word's CWN synset is the subordinate of another's); meronymy (i.e. one word's CWN synset is a part, member or substance of another's); nearsynonymy (i.e. words sharing in relatedness rather than within the same synset).

3.1. Data Collection

In order to include only prototypical pairs in EVALution-MAN, we filtered the CWN pairs and then further assessed them through manual annotation through both rating and tagging methods. The original number of word relation pairs from CWN stood at 50,000 entries. We excluded null or repeated word pairs, including instances of reversed order, i.e., 黑 (*hei*, black) vs. 白 (*bai*, white), and 白 (*bai*, white) vs. 黑 (*hei*, black). This first step of exclusion gave us a total of 10,000 word pairs. In order to maintain a balanced distribution of relata across the five relations types (snyonymy, antonymy, hyponymy, meronymy, and nearsynonymy) we included words that had a minimum of four other related words. This brought our total to 492 word pairs (376 relata).

We accordingly found that numerous pairs were not considered appropriate by our raters. For example, 鐵娘子 (*tie niangzi*, iron lady) vs. 鐵 (*tie*, iron) was not seen as related by our raters, yet seen as a hyponym pair in CWN; 如果说 (*ruguoshuo*: if we say), a modern variant of 果 (*guo*: if so), were deemed unrelated by our raters, yet seen a synonyms by CWN; 蘇 (*su*), a family name, was viewed as having a synonymy relation by CWN with 蘇東坡 (*Su Dongpo*) [3], however our raters rejected such a relation.

Another relevant issue we encountered was related to variances in Chinese, and in particular to its geography. Sinica Corpus – which is the base of CWN – was constructed from Taiwanese Mandarin language sources. Thus, variances in terminology such as, 北市 (*beishi*, North city) vs. 台北市 (*taibei shi*, Taipei city) did not fit the background of our raters, whom were all from Mainland China.

3.2. Reliability Check

The next step in constructing the dataset involved the rating of relatedness between word pairs. We divided the 494 word relation pairs extracted from CWN into two groups equalling 250 and 244 word relation pairs. Each word relation pair was placed in carrier sentences that represented a specific relation type. For example, while 龍和鳳相關 (*long he feng xiangguan*, "Dragon and phoenix are related to each other.") represented the relation type of near synonymy, 兔子和兔的意思相近 (*tuzi he tu de yisi xiangjin*, "兔子 (*tuzi*, rabbit) is simillar to 兔 (*tu*, rabbit)") represented the relation type of synonymy. Two documents were then constructed from the two groups of statements, each alongside a rating criterion: 'totally agree', 'agree', 'don't know', 'don't agree'and 'totally disagree'. Ten linguistics Ph.D. students (5 for each list of statements) were asked to rate the statements according to the abovementioned criterion. We also added 'don't know X'and 'don't know Y'tags in case a rater was not familiar with a given word.

Only pairs that had at least three positive ratings ("totally agree", and "agree") were included into the positive results. The rest were labeled as negative results. The results of the rating procedure revealed that of the initial 492 word pairs there were 360 positive word pairs and 132 negative pairs. Table 1 details the number of positive pairs and relata per relation type. Note that the number of relata (494) is the total number across all relation types. After extracting all the words in the positive pairs (hence, the relata and relatum), and filtering out repeated words, we arrived at 376 relata from the 360 positive word pairs.

Table 1: Numbers of Checking results

Relation	Pairs	Relata
Synonymy	61	114
Antonymy	34	50
Hyponymy	185	247
Meronymy	19	32
Nearsynonymy	61	51
Total	360	494

3.3. Semantic Tagging

As a further step in describing the dataset, we identified the semantic type information of the positive pairs. For the 376 relata, their total frequency and PoS distribution were calculated in a combined corpus that included Sinica Corpus (Huang et al., 2004) and Chinese GigaWord (Huang et al., 2010).

An additional three Ph.D. students were then asked to tag these relata according to their semantic types:

1. Basic/Subordinate/Superordinate: 研討會 (*yantaohui*, seminar) can be tagged as "Basic" while 會議 (*huiyi*, meeting) can be tagged as "Superodinate";
2. General/Specific: 台北 (*Taipei*, Taipei) can be tagged as "Specific" while 城市 (*chengshi*, city) can be tagged as "General";
3. Abstract/Concrete: 玫瑰 (*meigui*, rose) can be tagged as "Concrete" while 概念 (*gainian*, concept) can be tagged as "Abstract";
4. Event/Action/Time/Space/Object/Animal /Plant/Food/Color/People/Attribute: such as 轉變成 (*zhuanbiancheng*, transform into) can be tagged as "Action".

Only relata that showed agreement between at least two of the taggers were treated as positive results, leaving the total

[3] 蘇東坡 (*Su Dongpo*) is the name of a Chinese poet of the Song Dynasty

set of semantically tagged relata at 373. Meanwhile, frequency information was calculated in a combined corpus of Chen et al. (1996) and Hong and Huang (2006).

4. Conclusion and Future Work

EVALution-MAN is a dataset of Mandarin word relation pairs in Traditional Chinese for training and evaluation of DSMs or other applications. It has been manually rated according to relation pairs, and tagged for semantic type by native Mandarin speakers. It is freely available online at "https://github.com/LHongchao/EVALution_MAN". Future work will focus on extending the number of manually tagged relation pairs through extracting existing pairs through other ontology resources and/or through the use of behavioral methods that elicit semantic types of given words. A second goal of future work will be to provide a dataset for Mandarin in simplified Chinese.

5. Acknowledgements

This work is partially funded by the Hong Kong PhD Fellowship Scheme for Enrico Santus under PF12-13656.

6. Bibliographical References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics.

Benotto, G. (2015). Distributional models for semantic relations: A sudy on hyponymy and antonymy.

Chen, K.-J., Huang, C.-R., Chang, L.-P., and Hsu, H.-L. (1996). Sinica corpus: Design methodology for balanced corpora. *Language*, 167:176.

Fellbaum, C. (1998). *WordNet*. Wiley Online Library.

Girju, R., Badulescu, A., and Moldovan, D. (2006). Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1):83–135.

Hamp, B., Feldweg, H., et al. (1997). Germanet-a lexical-semantic net for german. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15. Citeseer.

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

Hill, F., Reichart, R., and Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Hong, J.-F. and Huang, C.-R. (2006). Using chinese gigaword corpus and chinese word sketch in linguistic research. In *The 20th Pacific Asia Conference on Language, Information and Computation (PACLIC-20). November*, pages 1–3.

Huang, C.-R., Chang, R.-Y., and Lee, H.-P. (2004). Sinica bow (bilingual ontological wordnet): Integration of bilingual wordnet and sumo. In *LREC*.

Huang, J., Hsieh, S.-K., Hong, J.-F., Chen, Y.-Z., Su, I.-L., Chen, Y.-X., and Huang, S.-W. (2010). Chinese wordnet: Design, implementation, and application of an infrastructure for cross-lingual knowledge processing. *Journal of Chinese Information Processing*, 24(2):14–23.

Jia, Z., He, D., Yin, H., and Li, T. (2014). Acquisition of part-whole relations based on unsupervised learning. *Journal of Southwest Jiaotong University*, 49(4):590–596.

Landauer, T. K. and Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Liu, Q. and Li, S. (2002). Word similarity computing based on how-net. *Computational Linguistics and Chinese Language Processing*, 7(2):59–76.

Mei, J., Zhu, Y., Gao, Y., and Yin, H. (1983). *Cilin-Chinese thesaurus*. Shanghai Lexicographical Publishing House.

Mohammad, S., Dorr, B., and Hirst, G. (2008). Computing word-pair antonymy. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 982–991. Association for Computational Linguistics.

Murphy, M. L. (2003). *Semantic relations and the lexicon: antonymy, synonymy and other paradigms*. Cambridge University Press.

Pantel, P. and Pennacchiotti, M. (2006). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 113–120. Association for Computational Linguistics.

Paradis, C., Willners, C., and Jones, S. (2009). Good and bad opposites using textual and experimental techniques to measure antonym canonicity. *The Mental Lexicon*, 4(3):380–429.

Santus, E., Lenci, A., Lu, Q., and Im Walde, S. S. (2014a). Chasing hypernyms in vector spaces with entropy. In *EACL*, pages 38–42.

Santus, E., Lu, Q., Lenci, A., and Huang, C.-R. (2014b). Taking antonymy mask off in vector space. In *Proceedings of PACLIC*, pages 135–144.

Santus, E., Lu, Q., Lenci, A., and Huang, C. (2014c). Unsupervised antonym-synonym discrimination in vector space.

Santus, E., Yung, F., Lenci, A., and Huang, C.-R. (2015). Evalution 1.0: an evolving semantic dataset for training and evaluation of distributional semantic models. *ACL-IJCNLP 2015*, page 64.

Sun, S. (2014). *Research on statistical word-level semantic relatedness computation*. Ph.D. thesis, Harbin Institute of Technology.

Van Hage, W. R., Kolb, H., and Schreiber, G. (2006). A method for learning part-whole relations. In *The Semantic Web-ISWC 2006*, pages 723–735. Springer.

Wang, B. (1999). *Automatic Chinese-English Paragraph Segmentation and Alignment*. Ph.D. thesis, The Chinese Academy of Sciences.

7. Language Resource References

Baroni, M. and Lenci, A. (2011). How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10. Association for Computational Linguistics.

Benotto, G. (2015). Distributional models for semantic relations: A sudy on hyponymy and antonymy.

Huang, J., Hsieh, S.-K., Hong, J.-F., Chen, Y.-Z., Su, I.-L., Chen, Y.-X., and Huang, S.-W. (2010). Chinese wordnet: Design, implementation, and application of an infrastructure for cross-lingual knowledge processing. *Journal of Chinese Information Processing*, 24(2):14–23.

Landauer, T. K. and Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Liu, Q. and Li, S. (2002). Word similarity computing based on how-net. *Computational Linguistics and Chinese Language Processing*, 7(2):59–76.

Santus, E., Yung, F., Lenci, A., and Huang, C.-R. (2015). Evalution 1.0: an evolving semantic dataset for training and evaluation of distributional semantic models. *ACL-IJCNLP 2015*, page 64.