VCU at Semeval-2016 Task 14: Evaluating similarity measures for semantic taxonomy enrichment

Bridget T. McInnes
Department of Computer Science
Virginia Commonwealth University
Richmond, VA 23284
btmcinnes@vcu.edu

Abstract

This paper describes the VCU systems that participated in the Semantic Taxonomy Enrichment task of SemEval 2016. The three systems are unsupervised and relied on dictionary-based similarity measures. The first two runs used first-order measures (Lesk and First-order vector), and the third run used a second-order measure (Second-order vector). The first-order measures obtained a higher Wu & Palmer score than the second-order measure on the test data. All three runs obtained a higher Wu & Palmer and F1 score than the random baseline but not the first-word first-sense baseline.

1 Introduction

Semantic knowledge bases are used in a number of NLP applications, e.g. Word Sense Disambiguation ((Agirre et al., 2014) (Agirre et al., 2010)) and Information Retrieval (Uddin et al., 2013). These knowledge bases span across a number of domains, e.g. WordNet, Gene Ontology, and Medical Subject Headings. Building and maintaining these knowledge bases is expensive and manually intensive (Martinez-Gil, 2015). Semantic taxonomy enrichment aims to aid in the maintenance process by automatically placing new terms into an existing taxonomy.

The Semantic Taxonomy Enrichment task (SemEval 2016 Task 14) objective was to automatically incorporate new word senses into WordNet [1], a lexical dictionary of English terms that are linked together based on a number of relations (e.g. is-a). In this task, systems were provided a list of $out\text{-}of\text{-}vocabulary$ terms (OOVs), their part-of-speech (POS), and a brief description of the term (OOV description). The goal of this task was to automatically identify which WordNet synset the OOV should be associated with and whether the association was synonymous ($merge$) or the OOV is a hyponym of the synset ($attach$).

Semantic similarity measures have been shown useful in the development of terminologies and ontologies (Vizenor et al., 2009). These measures quantify how related two terms are. The measures that we focus on for this task rely on definitional information extracted from knowledge sources such as WordNet. Our approach to the shared task is a modification of these measures relying on WordNet's gloss information, and the OOV descriptions.

We categorize the measures we evaluated as first-order and second-order measures. The first-order measures conduct a direct comparison between the words in the WordNet synset's gloss and the words in the OOV description. Second-order measures incorporate additional context by creating a vector for each word in the synset's gloss (or OOV description) containing words that co-occur with it from an external corpus. These word vectors are then averaged to create a single co-occurrence vector for the OOV or synset.

The VCU systems were implemented using the freely available, open source UMLS-Similarity package (McInnes et al., 2009) (version 1.45), which includes support for user-defined dictionaries and corpora, in addition to the first-order and second-order measures.

[1] https://wordnet.princeton.edu/

Proceedings of SemEval-2016, pages 1351–1355,
San Diego, California, June 16-17, 2016.

The paper is organized as follows. First, we describe the details of the three VCU systems that participated in this task. Second, we discuss the evaluation metrics. Lastly, we discuss the results of the systems.

2 VCU Systems

There were three VCU systems. VCU-Run-1 uses the $Lesk$ measure; VCU-Run-2 uses the $First\text{-}order\ vector$ measure; and VCU-Run-3 uses the $Second\text{-}order\ vector$ measure. Our goal was to compare the different measures on the task of identifying the appropriate placement for an OOV in WordNet. The measure used to identify the appropriate WordNet synset for a given OOV is the only difference between the three runs.

In the VCU systems, an OOV, its POS and text description is taken as input. First, the WordNet synsets are filtered based on their POS (see POS Filtering). Second, a score is assigned to each WordNet synset (see Measures). Third, the synset with the highest score is assigned to the OOV; if the score is greater than 0.7 it is labeled $merge$ (the two terms are synonymous), otherwise $attach$ (the OVV is a hyponym of the synset). The 0.7 score was set at development time and more work is required to determine what threshold should be used.

2.1 Measures

The VCU systems explored using three dictionary-based measures to identify the degree to which the OOV was similar to a WordNet synset for placement in the taxonomy. The first-order measure referred to as Lesk and First-order vector; and the third is a second-order measure referred to as Second-order vector. This subsection describes these measures.

2.1.1 Lesk

The Lesk measure, initially proposed by (Lesk, 1986), quantifies the relatedness between two terms by counting the number of overlaps between their two definitions. An overlap is defined as the longest sequence of one or more consecutive words that occur in both definitions. The length of the overlap is squared to give a greater weight to longer overlaps. For example, given the definitions *a very large number* and *a very large indefinite number*. The overlap *very large* would be given the score 4, the overlap *number* would be given a score of 1, and the total Lesk score would be 5.

2.1.2 First-order vector

First-order vector is a modification of the Lesk measure. It treats each word in the definition as an element in a vector. A vector is created for the WordNet synset and the OOV where the element in the vector is the number of times the associated word occurred in their respective definitions. The cosine similarity is then used to quantify the degree to which the two terms are similar. Figure 1 shows a simple example where the OOV definition is *a very large number* and the WordNet definition is *a very large indefinite number*.

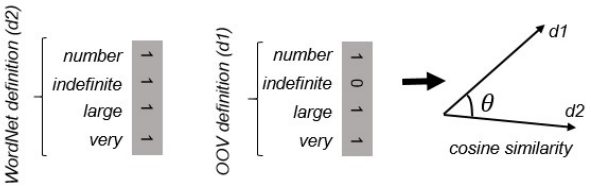

Figure 1: First-order vector example

2.1.3 Second-order vector

One of the disadvantages to first-order measures is that they rely on the exact matching of the words in the definitions. Therefore, *a humongous sum* would obtain a Lesk and First-order vector score of zero when compared to a *a very large number* even though both descriptions are clearly associated. The Second-order vector measure was introduced by (Patwardhan and Pedersen, 2006) to alleviate this.

In this measure, a vector is created for each word in the definition containing the words that co-occur with words from an external corpus. These word vectors are averaged to create a single co-occurrence vector for the OOV or synset. The similarity between the OOV and synset is then calculated by taking the cosine between their respective second-order vectors. Figure 2 shows a simple example using the "very large number" example again from above.

2.2 Definition Creation

The $definitions$ we use in the VCU systems consist of: 1) the OOV description and 2) the WordNet synset's gloss.

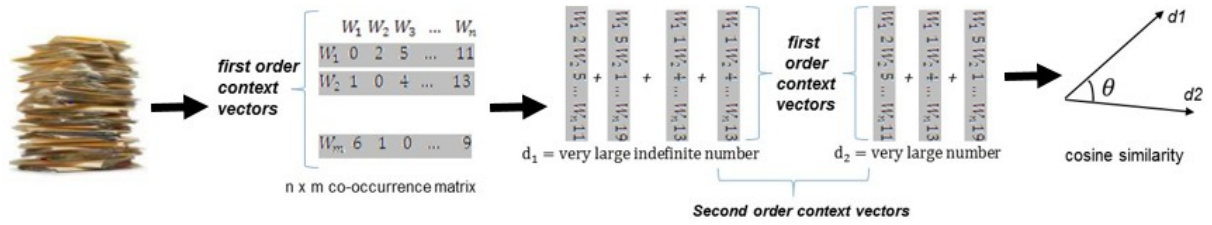

Figure 2: Second-order vector example

When implementing Lesk in WordNet, (Banerjee and Pedersen, 2002) found that the WordNet glosses were short, and did not contain enough overlaps to distinguish between multiple synsets, therefore, they extended this measure to include the glosses of the related concepts. This modification has been shown to improve the performance of both Lesk and Second-order vectors on the tasks of Word Sense Disambiguation (Banerjee and Pedersen, 2002) and Semantic Similarity (Patwardhan, 2003) (McInnes and Pedersen, 2013). The VCU Systems included the glosses of the following WordNet relations: holonyms ($holo$), hypernyms ($hype$), hyponyms ($hypo$), and meronyms ($mero$). Therefore, our WordNet synset $definition$ contains not only the gloss of the synset but also the glosses of the related synsets.

The stop words were removed from the OOV descriptions and WordNet glosses prior to processing using the stoplist from the Ngram Statistics Package [2] (NSP), and all of the words were lower cased. No additional pre-processing (e.g. stemming) was conducted.

2.3 Matrix for second-order measure

In the VCU system, the second-order co-occurrence matrix was created using bigram counts obtained from the GigaWord Corpus 2nd Edition Agence France-Presse, English Service (afp_eng) data. We selected only those bigrams based on the following criteria: (1) neither word in the bigram was a stop word in the NSP stoplist; and (2) the bigram occurred at least twice in the corpus. The Ngram Statistics Package (Banerjee and Pedersen, 2003) was used to collect the bigram and their frequency counts.

[2]https://sourceforge.net/projects/ngram/

The resulting matrix contained 59,217 rows and 70,726 columns, representing 4,487,663 unique bigrams. The matrix is not symmetric because the co-occurrences are bigrams. Therefore, the number of times *school bus* was seen in the text is different than the number of times *bus school*. All the words in the corpus were lower cased prior to processing.

2.4 Part-of-Speech Filters

Due to the number of synsets in WordNet, we filtered them based on two criteria: (1) the WordNet synset has the same POS as the OOV; and (2) the WordNet synset contained a word of the same POS as a word in the OOV description. These criteria were adaptations of the first-word/first-sense baseline that was provided by the organizers to reduce the number of possible synset choices. To obtain the POS of the words in the OOV descriptions, we used the OpenNLP POS Tagger (Baldridge et al., 2002).

3 Evaluation Metrics

The VCU system was evaluated using four metrics: WuP, Lemma Match, Recall and F1. WuP is based on the similarity metric proposed by (Wu and Palmer, 1994). It was used to evaluate how close the system came to identifying the correct synset. In this metric, the similarity is twice the depth of the two synsets ($synset_{sys}$ and $synset_{gold}$) Least Common Subsumer (LCS) divided by the sum of the depths of the individual synsets as defined in Equation 1. The LCS is the most specific ancestor shared by two synsets using WordNet's hypernymy/hyponymy relations.

$$\text{WuP} = \frac{2 * \text{depth}(\text{lcs}(synset_{sys}, synset_{gold}))}{\text{depth}(synset_{sys}) + \text{depth}(synst_{gold})} \quad (1)$$

System	Measure	WuP	Lemma Match	Recall	F1
VCU-Run-1	First-order Lesk	0.4190	0.1706	0.9967	0.5900
VCU-Run-2	First-order vector	0.4317	0.1605	0.9967	0.6024
VCU-Run-3	Second-order vector	0.4076	0.1237	0.9967	0.5786
Baseline	Random	0.2269	0.0000	1.000	0.3699
Baseline	First word/sense	0.5134	0.4150	1.000	0.6790

Table 1: VCU System Test Results

Recall determines how many of the OOV terms were assigned to a WordNet synset. The F1 score is the F-measure (harmonic mean) of WuP and Recall. The Lemma Match is the percentage of system synsets that exactly matched the gold standard.

4 Results and Discussion

Table 1 shows the VCU system results for the Semantic Taxonomy Enrichment task, and the two baselines provided by the organizers (Random and First-word First-sense). The Random baseline chooses a random synset of the appropriate POS. The First-word First-sense (First word/sense) assigns the first synset whose word is also the first head word in the OOV description with the same POS.

The Recall results show that the VCU systems did not assign all of the OOVs to a WordNet synset. Investigation into this found that the POS filter (see Section 2.4) was too aggressive and no WordNet synset met the filtering criteria. The WuP and Lemma Match results show that the VCU systems obtained a higher score than the random baseline but not the First word/sense baseline.

The results between the three VCU system runs show that the First-order measures obtained a higher Lemma Match, WuP and F1 score than the Second-order vector measure. This indicates that the additional contextual information from the corpus did not provide useful information, hurting the performance. The co-occurrence matrix created for the Second-order vector measure was based on text from the GigaWord corpus. Looking back, we believe that utilizing a more up-to-date text such as Wikipedia, or using WordNet as a corpus itself as (Pedersen, 2014) may increase the performance the results.

The results between the first-order measures show Lesk obtained a lower WuP score than First-order vector, but a higher Lemma Match. Analysis of the mappings found that Lesk was merging all of the OOVs to the WordNet synsets rather than attaching them as a hyponym. When assigning $attach$ rather than $merge$ to each of the mappings, the WuP score for Lesk increased to 0.4461 (higher than First-order vector's 0.4317). This indicates that additional work is required to determine whether the attachment of the OOV to the synset should be a $merge$ or an $attach$.

5 Conclusion and Future Work

In this paper, we described the VCU systems that participated in the Semantic Taxonomy Enrichment task of SemEval 2016. The three systems are unsupervised and relied on dictionary-based similarity measures. The first two runs used First-order measures (Lesk and First-order vector), and the third run used a second-order measure (Second-order vector). The first-order measures obtained a higher Wu & Palmer score than the second-order measure.

Analysis of the OOV mappings to WordNet synsets by the measures highlighted three areas of future work. First, more attention needs to be paid in determining whether the new term should be merged as a synonym or attached as a hyponym to the synset. Second, a more indepth analysis of the definitions used to represent the context of the new terms and the WordNet synsets needs to be conducted. Lastly, although the Second-order vector did not perform as well as the First-order measures, the type of the corpus used to create the second-order vectors needs to be explored.

Acknowledgments

WordNet was accessed using WordNet–QueryData [3]. The VCU systems were implemented using UMLS-Similarity [4]. I would like to thank Ted Pedersen for help in understanding the task, and useful brainstorming discussions.

[3] http://search.cpan.org/WordNet-QueryData

[4] http://search.cpan.org/UMLS-Similarity

References

E. Agirre, A. Soroa, and M. Stevenson. 2010. Graph-based word sense disambiguation of biomedical documents. *Bioinformatics*, 26(22):2889–2896.

E. Agirre, O.L. de Lacalle, and A. Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.

J. Baldridge, T. Morton, and G. Bierner. 2002. The opennlp maximum entropy package. *tech. rep., SourceForge*.

S. Banerjee and T. Pedersen. 2002. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In *Proceedings of the Third Intl. Conf. on Intell. Text Process. and Comp. Ling.*, pages 136–145, Mexico City, Mexico, February. Springer.

Satanjeev Banerjee and Ted Pedersen. 2003. The design, implementation, and use of the ngram statistics package. In *Computational Linguistics and Intelligent Text Processing*, pages 370–381. Springer.

O. Bodenreider and A. Burgun. 2004. Aligning knowledge sources in the UMLS: methods, quantitative results, and applications. In *Proceedings of the 11th World Congress on MEDINFO*, pages 327–331, San Fransico, CA, November.

M. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proc. of the 5th Annual Intl. Conf. on Sys. Doc.*, pages 24–26, Toronto, Canada, June.

J. Martinez-Gil. 2015. Automated knowledge base management: A survey. *Computer Science Review*, 18:1–9.

B.T. McInnes and T. Pedersen. 2013. Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. *Journal of Biomedical Informatics*, 46(6):1116–1124.

B.T. McInnes, T. Pedersen, and S.V. Pakhomov. 2009. UMLS-Interface and UMLS-Similarity: Open source software for measuring paths and semantic similarity. In *Proc. of the AMIA Symposium*, pages 431–435, San Fransico, CA, November.

S. Patwardhan and T. Pedersen. 2006. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together*, pages 1–8, Trento, Italy, April.

S. Patwardhan. 2003. Incorporating dictionary and corpus information into a context vector measure of semantic relatedness. *Master's thesis, University of Minnesota, Duluth*.

Ted Pedersen. 2014. Duluth: Measuring cross–level semantic similarity with first and second–order dictionary overlaps. *SemEval 2014*, page 247.

M.N. Uddin, T.H. Duong, N.T. Nguyen, X. Qi, and G.S. Jo. 2013. Semantic similarity measures for enhancing information retrieval in folksonomies. *Expert Systems with Applications*, 40(5):1645–1653.

Lowell T Vizenor, Olivier Bodenreider, and Alexa T McCray. 2009. Auditing associative relations across two knowledge sources. *Journal of biomedical informatics*, 42(3):426–439.

Z. Wu and M. Palmer. 1994. Verbs semantics and lexical selection. In *Proc. of the 32nd Meeting of ACL*, pages 133–138, Las Cruces, NM, June.