CFILT-CORE: Finding Semantic Textual Similarity using UNL

Avishek Dan
IIT Bombay
Mumbai, India
avishekdan@cse.iitb.ac.in

Pushpak Bhattacharyya
IIT Bombay
Mumbai, India
pb@cse.iitb.ac.in

Abstract

Semantic Textual Similarity is the task of finding the degree of similarity between a pair of sentences through semantics extraction. This is motivated by the fact that syntactically diverse sentences often convey the same meaning. This paper describes the approach that was used in the *SEM Shared Task 2013. The approach combines semantic, syntactic and lexical similarity measures for finding similarity scores between sentences. We describe a Universal Networking Language based semantic extraction system for measuring the semantic similarity.

1 Introduction

The core Semantic Textual Similarity shared task of *SEM 2013 (Agirre et al., 2013) is to generate a score in the range 0-5 for a pair of sentences depending on their semantic similarity. Universal Networking Language (UNL) (Uchida, 1996) is an ideal mechanism for meaning representation. Our system first converts the sentences into UNL graph representation and then matches the graphs to generate the degree of semantic relatedness. Even though the goal is to judge sentences based on their semantic relatedness, our system incorporates some lexical and syntactic similarity measures to make the system robust in the face of data sparsity.

The following sections give a brief introduction to UNL, decribe our English Enconverter, discuss the various similarity measures used in the task, the training of the system and the results obtained on the task datasets.

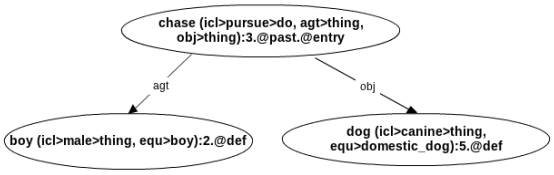

Figure 1: UNL Graph for 'The boy chased the dog'

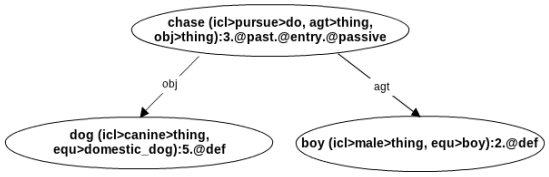

Figure 2: UNL Graph for 'The dog was chased by the boy'

2 Universal Networking Language

Universal Networking Language (UNL) is an interlingua that represents a sentence in a language independent, unambiguous form. UNL representations have a graphical structure with concepts being represented as nodes and relations between concepts being represented by edges between the nodes. Figure 1 shows the UNL graph corresponding to the sentence 'The boy chased the dog.' The conversion from a source language to UNL is called enconversion. The three main building blocks of UNL are relations, universal words and attributes.

2.1 Universal Words

Universal words (UWs) are language independent concepts that are linked to various language resources. The UWs used by us are linked to the

Princeton wordnet and various other language wordnet synsets. UWs consist of a head word which is the word in its lemma form. This is followed by a constraint list that is used to disambiguate it. For example, chase icl (includes) pursue indicates that chase as an act of pursuing is indicated here.

2.2 Relations

Relations are two place functions that imdicate the relationship between UWs. Some of the commonly used relations are agent (agt), object (obj), instrument (ins), place (plc). For example, in figure 1 the relation agt between boy and chase indicates that the boy is the doer of the action.

2.3 Attribute

Attributes are one place functions that convey various morphological and pragmatic information. For example, in figure 1 the attribute @past indicates that the verb is in the past tense.

2.4 Enconversion

The conversion from English to UNL involves augmenting the sentence with various factors such as POS tags, NER tags, dependency parse tree paths. The suitable UW generation is achieved through a word sense disambiguation module. The attribute and relation generation is achieved through a combination of rule-base and classifiers trained on a small corpus.

The syntax independent structure of UNL makes it very suitable for the similarity task. For example, if the example in figure 1 is passivized, the UNL graph structure remains essentially the same with only an additional attribute passive indicating the voice as indicated in figure 2.

3 Similarity Measures

We broadly define three categories of similarity measures based on our classification of perception of similarity.

3.1 Lexical Similarity Measures

Lexical similarity measures consider the sentences as set-of-words. These measures are motivated by our view that sentences having a lot of common words will appear quite similar to a human user. For computing the lexical similarity measures, the sentence is tokenized using Stanford Parser and a set is created out of the generated tokens.

3.1.1 Jaccard Similarity Coefficient

The Jaccard coefficient compares the similarity or diversity of two sets. It is the ratio of size of intersection to the size of union of the sets.

$$JSim(S1, S2) = \frac{|S1 \cap S2|}{|S1 \cup S2|}$$

3.1.2 Extended Jaccard Similarity Coefficient

We define a new measure based on the Jaccard similarity coefficient that captures the relatedness between words. The tokens in the set are augmented with related words from Princeton Wordnet. As a preprocessing step, all the tokens are stemmed using Wordnet Stemmer. For each possible sense of each token, its synonyms, antonyms, hypernyms and holonyms are added to the set as applicable. For example, hypernyms are added only when the token appeared as a noun or verb in the Wordnet. The scoring function used is defined as

$$ExtJSim(S1, S2) = \frac{|ExtS1 \cap ExtS2|}{|S1 \cup S2|}$$

The following example illustrates the intuition behind this similarity measure.

- I am cooking chicken in the house.
- I am grilling chicken in the kitchen.

The measure generates a similarity score of 1 since grilling is a kind of cooking (hypernymy) and kitchen is a part of house (holonymy).

3.2 Syntactic Similarity Measures

Structural similarity as an indicator of textual similarity is captured by the syntactic similarity measures. Parses are obtained for the pair of English sentences using Stanford Parser. The parser is run on the English PCFG model. The dependency graphs of the two sentences are matched to generate the similarity score. A dependency graph consists of a number of dependency relations of the form *dep(word1, word2)* where dep is the type of relation and word1 and word2 are the words between which the relation holds. A complete match of a dependency relation contributes 1 to the score whereas a match of

only the words in the relation contributes 0.75 to the score.

$$SynSim(S1, S2) = \frac{|S1 \cap S2|}{|S1 \cup S2|} + 0.75* \frac{\sum_{a \in S1, b \in S2}[[a.w1 = b.w1 \& a.w2 = b.w2]]}{|S1 \cup S2|}$$

Here S1 and S2 represent the set of dependency relations.

An extended syntactic similarity measure in which exact word matchings are replaced by a match within a set formed by extending the word with related words as described in 3.1.2 is also used.

3.3 Semantic Similarity Measure

Semantic similarity measures try to capture the similarity in the meaning of the sentences. The UNL graphs generated for the two sentences are compared using the formula given below. It is of note that in this case we consider a combined formula incorporating the basic as well as Wordnet extended cases. In addition, synonymy is no more used for enriching the word bank since UWs by design are mapped to synsets, hence all synonyms are equivalent in a UNL graph.

$$SemSim(S1, S2) = \frac{|S1 \cap S2|}{|S1 \cup S2|} + \sum_{a \in S1, b \in S2} (0.75* \frac{[[a.w1 = b.w1 \& a.w2 = b.w2]]}{|S1 \cup S2|} + 0.75* \frac{[[a.r = b.r \& a.Ew1 = b.Ew1 \& a.Ew2 = b.Ew2]]}{|S1 \cup S2|} + 0.6 * \frac{[[a.Ew1 = b.Ew1 \& a.Ew2 = b.Ew2]]}{|S1 \cup S2|})$$

4 Corpus

The system is trained on the Semantic Textual Similarity 2012 task data. The training dataset consists of 750 pairs from the MSR-Paraphrase corpus, 750 sentences from the MSR-Video corpus and 734 pairs from the SMTeuroparl corpus.

The test set contains headlines mined from several news sources mined by European Media Monitor, sense definitions from WordNet and OntoNotes, sense definitions from WordNet and FrameNet, sentences from DARPA GALE HTER and HyTER, where one sentence is a MT output and the other is a reference translation.

Each corpus contains pairs of sentences with an associated score from 0 to 5. The scores are given based on whether the sentences are on different topics (0), on the same topic but have different content (1), not equivalent but sharing some details (2), roughly equivalent with some inportant information missing or differing (3), mostly important while differing in some unimportant details (4) or completely equivalent (5).

5 Training

The several scores are combined by training a Linear Regression model. We use the inbuilt libaries of Weka to learn the weights. To compute the probability of a test sentence pair, the following formula is used.

$$score(S1, S2) = \sum_{i=1}^{5} \lambda_i score_i(S1, S2)$$

6 Results

The test dataset contained many very long sentences which could not be parsed by the Stanford parser used by the UNL system. Hence erroneous output were produced in these cases. Table 1 summarizes the results.

The UNL system is not robust enough to handle large sentences with long distance relationships which leads to poor performance on the OnWN and FNWN datasets.

7 Conclusion and Future Work

The approach discussed in the paper shows promise for the small sentences. The ongoing development of UNL is expected to improve the accuracy of the system. Tuning the scoring parameters on a development set instead of arbitrary values may improve results.

References

Eneko Agirre and Daniel Cer and Mona Diab and Aitor Gonzalez-Agirre and Weiwei Guo. *SEM 2013 Shared Task: Semantic Textual Similarity, including a Pilot on Typed-Similarity. *SEM 2013: The Second

Table 1: Results

Corpus	CFILT	Best Results
Headlines	0.5336	0.7642
OnWN	0.2381	0.7529
FNWN	0.2261	0.5818
SMT	0.2906	0.3804
Mean	0.3531	0.6181

Joint Conference on Lexical and Computational Semantics. Association for Computational Linguistics.

Hiroshi Uchida. UNL: Universal Networking LanguageAn Electronic Language for Communication, Understanding, and Collaboration. 1996. UNU/IAS/UNL Center, Tokyo.