CFILT-CORE: Semantic Textual Similarity using Universal Networking Language

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

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

Abstract

This paper describes the system that was submitted in the *SEM 2013 Semantic Textual Similarity shared task. The task aims to find the similarity score between a pair of sentences. We describe a Universal Networking Language (UNL) based semantic extraction system for measuring the semantic similarity. Our approach combines syntactic and word level similarity measures along with the UNL based semantic similarity measures for finding similarity scores between sentences.

1 Introduction

Semantic Textual Similarity is the task of finding the degree of semantic equivalence between a pair of sentences. 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. Textual similarity finds applications in information retrieval and it is closely related to textual entailment. Universal Networking Language (UNL) (Uchida, 1996) is an ideal mechanism for semantics representation. Our system first converts the sentences into a UNL graph representation and then matches the graphs to generate the semantic relatedness score. 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.

Section 2 give a brief introduction to UNL. Section 3 decribes the English Enconverter developed by us. Section 4 discusses the various similarity measures used for the task. Section 5 mentions the corpus used for training and testing. Section 6 describes the method used to train the system and Section 7 presents the results obtained on the task datasets.

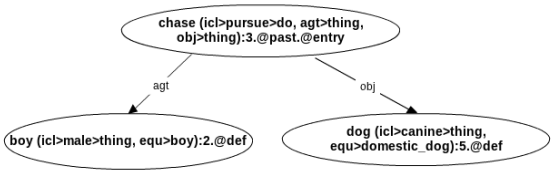

Figure 1: UNL Graph for *'The boy chased the dog'*

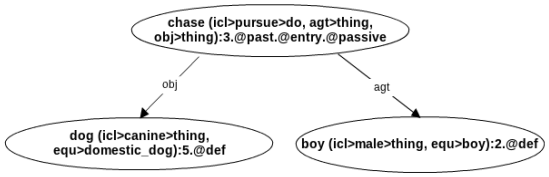

Figure 2: UNL Graph for *'The dog was chased by the boy'*

2 Universal Networking Language

Universal Networking Language (UNL) is an interlingua that represents a sentence in a language independent, unambiguous form. The three main building blocks of UNL are relations, universal words and attributes. UNL representations have a graphical structure with concepts being represented as nodes (universal words) and interactions between concepts being represented by edges (relations) between the nodes. Figure 1 shows the UNL graph correspond-

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 216–220, Atlanta, Georgia, June 13-14, 2013.

ing to the sentence *'The boy chased the dog.'* The conversion from a source language to UNL is called enconversion. The reverse process of generating a natural language sentence from UNL is called deconversion. The enconversion process is markedly more difficult than the deconversion process due to the inherent ambiguity and idiosyncracy of natural language.

UNL representation captures the semantics independent of the structure of the language. Figures 1 and 2 show the UNL representation of two structurally different sentences which convey the same meaning. The UNL graph structure remains the same with an additional attribute on the main verb of figure 2 indicating the voice of the sentence.

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. For example, in figure 2 the word *chased* is shown in its lemma form as *chased*. The head word is followed by a constraint list which is used to disambiguate it. For example, *chase* icl (includes) *pursue* indicates that chase as a type of pursuing is indicated here. Complex concepts are represented by hypernodes, which are UNL graphs themselves.

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.

3 UNL Generation

The conversion from English to UNL involves augmenting the sentence with various factors such as POS tags, NER tags and dependency parse tree relations and paths. The suitable UW generation is achieved through a word sense disambiguation (WSD) system trained on a tourism corpus. The WSD system maps the words to Wordnet 2.1 synset ids. The attribute and relation generation is achieved through a combination of rule-base and classifiers trained on a small corpus. We use a nearest neighbor classifier trained on the EOLSS corpus for generating relations. The attributes are generated by conditional random fields trained on the IGLU corpus. The attribute generation is a word level phenomena, hence attributes for complex UWs cannot be generated by the classifiers. The steps are described in detail.

3.1 Parts of Speech Tagging

The Stanford POS tagger using the WSJ corpus trained PCFG model is used to tag the sentences. Penn Treebank style tags are generated.

3.2 Word Sense Disambiguation

A Supervised Word Sense Disambiguation (WSD) tool trained in Tourism domain is used. The WSD system takes a sequence of tagged words and provides the WordNet synset ids of all nouns, verbs, adjectives and adverbs in the sequence. The accuracy of the system is depends on the length of the input sentence.

3.3 Named Entity Recognition

Stanford Named Entity Recognizer is used to tag the words in the sentence. The tags may be PERSON, LOCATION or ORGANIZATION.

3.4 Parsing and Clause Marking

Stanford Parser is used to parse the sentences. Rules based on the constituency parse are used to identify the clause boundaries. The dependency parse is used for clause type detection. It is also used in the later stages of UNL generation.

The clauses are converted into separate simple sentences for further processing. Independent clauses can be trivially separated since they have complete sentential structure of their own. Dependent clauses are converted into complete sentences using rules based on the type of clause. For example, for the sentence, *That he is a good sprinter, is*

known to all, containing a nominal clause, the simple sentences obtained are *he is a good sprinter* and *it is known to all*. Here the dependent clause is replaced by the anaphora *it* to generate the sentence corresponding to the main clause.

3.5 UW Generation

WordNet synset ids obtained from the WSD system and the parts of speech tags are used to generate the UWs. The head word is the English sentence in its lemma form. The constraint list is generated from the WordNet depending on the POS tag.

3.6 Relation Generation

Relations are generated by a combination of rule base and corpus based techniques. Rules are written using parts of speech tags, named entity tags and parse dependency relations. The corpus based techniques are used when insufficient rules exist for relation generation. We use a corpus of about 28000 sentences consisting of UNL graphs for WordNet glosses obtained from the UNDL foundation. This technique tries to find similar examples from the corpus and assigns the observed relation label to the new part of the sentence.

3.7 Attribute Generation

Attributes are a combination of morphological features and pragmatic information. Attribute generation can be considered to be a sequence labeling task on the words. A conditional random field trained on the corpus described in section 5.1 is used for attribute generation.

4 Similarity Measures

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

4.1 Word based Similarity Measure

Word based similarity measures consider the sentences as sets-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. The sentences are tokenized using Stanford Parser. The Jaccard coefficient (Agirre and Ghosh and Mooney, 2000) compares the similarity or diversity of two sets. It is the ratio of size of intersection to the size of union of two sets. 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. (Pedersen and Patwardhan and Michelizzi, 2004) As a preprocessing step, all the tokens are stemmed using WordNet Stemmer. For each possible sense of each stem, its synonyms, antonyms, hypernyms and holonyms are added to the set as applicable. For example, hypernyms are added only when the token appears 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).

4.2 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 4.1 is also used.

4.3 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. 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.

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

5 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).

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

6 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.

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

7 Results

The test dataset contained many very long sentences which could not be parsed by the Stanford parser used by the UNL system. In addition, the performance of the WSD system led to numerous false negatives. Hence erroneous output were produced in these cases. In these cases, the word based similarity measures somewhat stabilized the scores. 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.

8 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. A log-linear model instead of the linear combination of scores may capture the relationships between the scores in a better way.

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 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.

Alexander Strehl and Joydeep Ghosh and Raymond Mooney Impact of similarity measures on web-page clustering. 2000. Workshop on Artificial Intelligence for Web Search (AAAI 2000).

Ted Pedersen and Siddharth Patwardhan and Jason Michelizzi WordNet:: Similarity: measuring the relatedness of concepts. 2004. Demonstration Papers at HLT-NAACL 2004. Association for Computational Linguistics.