janardhan: Semantic Textual Similarity using Universal Networking Language graph matching

Janardhan Singh
IIT Bombay,
Mumbai, India
janardhan
@cse.iitb.ac.in

Arindam Bhattacharya
IIT Bombay,
Mumbai, India
arindamb
@cse.iitb.ac.in

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

Abstract

Sentences that are syntactically quite different can often have similar or same meaning. The SemEval 2012 task of Semantic Textual Similarity aims at finding the semantic similarity between two sentences. The semantic representation of Universal Networking Language (UNL), represents only the inherent meaning in a sentence without any syntactic details. Thus, comparing the UNL graphs of two sentences can give an insight into how semantically similar the two sentences are. This paper presents the UNL graph matching method for the Semantic Textual Similarity(STS) task.

1 Introduction

Universal Networking language (UNL) gives the semantic representation of sentences in a graphical form. By comparing the similarity of these graphs, we inherently compare only the semantic content of the two sentences, rather than comparing the similarities in the syntax. Thus, the UNL graph matching strategy is a natural choice for the Semantic Textual Similarity(STS) task of SemEval 2012. UNL graphs are also used in textual entailment and interlingua based machine translation tasks. We use the UNL enconverter system at: `http://www.cfilt.iitb.ac.in /UNL_enco` to generate the UNL graphs of the sentences. For the two graphs, generated from the two sentences, we give a similarity score by matching the two graphs.

In the following sections we describe UNL matching strategy. section 2 describes the UNL system and why this approach is useful, section 3 describes the matching algorithm, section 4 describes the challenges faced in this approach, section 5 gives the results and finally section 6 gives the conclusion and the future scope.

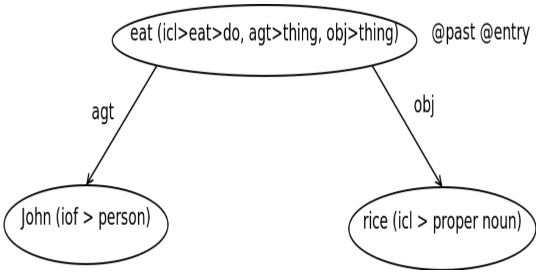

Figure 1: UNL graph for "John eats rice"

2 Universal Networking Language

The Universal Networking Language gives a graphical representation of the semantics of a text in the form of hypergraphs. The representation is at the semantic level which allows mapping of the similar meaning sentences having different syntax to the same representation. To exemplify this point, consider the UNL graphs generated for the following sentences:

Sentence 1: John ate rice.

Sentence 2: Rice was eaten by John.

The UNL graph generated from the system are given in figures 1 and 2 respectively.

The UNL graph consists of three components:

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 662–666,
Montréal, Canada, June 7-8, 2012. ©2012 Association for Computational Linguistics

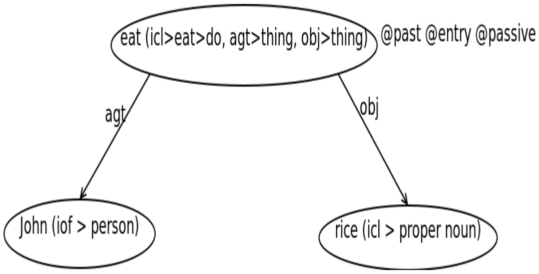

Figure 2: UNL graph for "Rice was eaten by John"

- Universal Words
- Relations
- Attributes

2.1 Universal Words

The Universal Words (UWs) form the vocabulary of the Universal Networking Language. They form the nodes of the UNL graph. The words are normalized to their basic lemma, for example, *eats* becomes *eat*. The Universal Word is, usually, followed by a disambiguating constraint list which is mainly used for disambiguating the sense of the Universal Word. For example, *John (iof > person)*, here the word *John* is disambiguated as an instance of (iof) a *person* and *rice* is disambiguated to be in the class of (icl) proper noun. The UNL generation system, uses a Universal word dictionary created using the wordnet.

2.2 Relations

The UNL manual describes 46 binary semantic relations among the Universal Words as given in UNL manual. These form the labelled arcs of the UNL graph. In the example of figures 1 and 2, the relations agent (agt) and object (obj) are shown. *John* is the *agent* of the action *eat* and *rice* is the object of the action *eat*. The UNL generation system generated these relations using complex rules based on the dependency and constituency parser outputs, Wordnet features and Named Entity recognizer output.

2.3 Attributes

Attributes are attached to the Universal Words to show the speakers perspective for some subjective information in the text. For the given example, with respect to the speaker of the text, the action of *eat* happened in the *past* with respect to the speaker. This is represented by the attribute @*past*.
The detailed description of the UNL standard can be found in the UNL manual available online at `http://www.undl.org/unlsys/unl/unl2005/`.

The two sentences listed above, have the same semantic content, although their syntax is different. One sentence is in the active voice, while the other sentence is in the passive. But if we compare the UNL graphs of the two sentences, they are almost identical, with an extra attribute @*passive* on the main verb *eat* in the second graph. The graph matching of the two sentences results in a high score near to 5. Like voice, most of the syntactic variations are dropped when we move from syntactic to semantic representation. Thus, comparing the semantic representation of the sentences, is useful, to identify their semantic similarity. The UNL generation system generates the attributes using similar features to those for relation generation.

3 UNL matching

The UNL system available online at: `http://www.cfilt.iitb.ac.in/UNL_enco`
produces graphs for the sentences by listing the binary relations present in the graph. An example of such a listing is :

Sentence 3: A man is eating a banana by a tree.

```
[unl:1]
agt ( eat(icl>eat>do, agt>thing,
obj>thing):4.@present.@progress
.@entry,
 man(icl>male>thing,
equ>adult_male):2.@indef )
ins ( eat(icl>eat>do, agt>thing,
obj>thing):4.@present.@progress
.@entry,
 tree(icl>woody_plant>thing)
:9.@indef )
obj ( eat(icl>eat>do, agt>thing,
obj>thing):4.@present.@progress
.@entry,
 banana(icl>herb>thing,
equ>banana_tree):6.@indef )
[\unl]
```

Sentence 4 : A man is eating a banana.

```
[unl:1]
agt ( eat(icl>eat>do, agt>thing,
obj>thing):4.@present.@progress
.@entry,
 man(icl>male>thing,
equ>adult_male):2.@indef )
obj ( eat(icl>eat>do, agt>thing,
obj>thing):4.@present.@progress
.@entry,
 banana(icl>herb>thing,
equ>banana_tree):6.@indef )
[\unl]
```

We treat the UNL graph of one sentence as *goldunl* and the other as *testunl*. The matching score between the two is found using the following formulation (Mohanty, 2008):

$$score(testunl, goldunl) = \frac{(2*precision*recall)}{(precision+recall)} \quad (1)$$

$$precision = \frac{\sum_{relation \in testunl} relation\_score(relation)}{(count(relations \in testunl))} \quad (2)$$

$$recall = \frac{\sum_{relation \in testunl} relation\_score(relation)}{(count(relations \in goldunl))} \quad (3)$$

$$relation\_score(relation) = avg(rel\_match, uw1score, uw2score) \quad (4)$$

$$rel\_match = \begin{cases} 1 & \text{if relation name matches} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

$$uwscore = avg(word\_score, attribute\_score) \quad (6)$$

$$word\_score = \begin{cases} 1 & \text{if universal word matches} \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

$$attribute\_score = F1score(testunl\_attr, goldunl\_attr) \quad (8)$$

The matching scheme is based on the idea of the F1 score. The two UNL graphs are a list of UNL relations each. Considering, one as the gold UNL graph and the other as the test UNL graph, we can find the precision and recall of the total relations that have matched. For the example given in section 2.4, the sentence 3 has three relations while sentence 4 has two relations. A correspondence between the relations *agt* of the two graphs and also the relation *obj* of the two graphs can be established based on the *universal words* that they connect. Each such relation match is given a score, explained later, which is used in the calculation of the precision and recall. From the precision and recall the F1 score can be easily calculated which becomes the total matching score of the two graphs.

The relation score is obtained by averaging the scores of relation match, and the score of the two universal word matches. The universal word match score has a component of the attributes that match between the corresponding universal words. This attribute matching is again the F1 score calculation similar to relation matching. Matching the attributes of the universal words, contributes to the score of the matched universal word, which in turn contributes to the score of the matched relation. Thus, matching of the semantic relations has more weight than the matching of the attributes.

The score obtained by this formulation is between 0 and 1. Another score between 0 and 1 is obtained by flipping the *goldunl* graph to *testunl* and *testunl* to *goldunl*. Average of these two scores is then multiplied by 5 to give the final score.

By this formulation, the score obtained by matching graphs for sentences 3 and 4 is 4.0

4 Challenges in the approach

In the UNL graph matching startegy we faced the following challenges:

4.1 Sentences with grammatical errors

Many of the sentences, especially, from the MSRpar dataset, had minor grammatical errors. The UNL generation requires grammatical correctness. Some of the examples of such sentences are:

- *The no-shows were Sens. John Kerry of Massachusetts and Bob Graham of Florida.*

- *She countersued for $125 million, saying G+J broke its contract with her by cutting her out of key editorial decisions and manipulated the magazine's financial figures.*

- *"She was crying and scared,' said Isa Yasin, the owner of the store.*

Here, terms like *G+J* and punctuation errors as in the third example lead to the generation of improper UNL graphs. To handle such cases, the UNL generation needs to get robust.

4.2 Scoping errors

UNL graphs are hypergraphs, in which, a node can in itself be a UNL graph. Scopes are given identity numbers like *:01,:02* and so on. While matching two different UNL graphs, this matching of scope identity numbers cannot be directly achieved. Also, one graph may have different number of scopes as compared to the other. Hence, eventhough the UNL graphs are generated correctly, due to scoping mismatches the matching score goes down. To tackle this problem, the UNL graphs generated are converted into scopeless form before the matching is performed. Every UNL graph has an entry node, which is the starting node of the graph. This is denoed by an *@entry* attribute on the node. Every scope, too, has an entry node. The idea for converting the UNL graphs into scopeless form is to replace the scope nodes by the graphs that these nodes represent, with the connection to the original scope node going to the entry node of the replacing graph.

4.3 Incomplete or no graph generation

It was observed that for some of the sentences, the UNL generation system did not produce UNL graphs or the generation was incomplete. Some of these sentences are:

- *The Metropolitan Transportation Authority was given two weeks to restore the $1.50 fare and the old commuter railroad rates, York declared.*

- *Long lines formed outside gas stations and people rushed to get money from cash machines Sunday as Israelis prepared to weather a strike that threatened to paralyze the country.*

These, are due to some internal system errors of the UNL generation system. To improve on this, the UNL generation system itself has to improve.

5 Results

By adopting the methodology described in section 3, the following results were obtained on the different datasets.

MSRpar	0.1936
MSRvid	0.5504
SMT-eur	0.3755
On-WN	0.2888
SMT-news	0.3387

As observed, the performance is good for the MSRvid dataset. This dataset consists of small and simple sentences which are grammatically correct. The performance on this dataset should further improve by capturing the synonyms of the Universal words while matching the UNL relations. The performance for MSRpar dataset is low. The sentences in this dataset are long and sometimes with minor grammatical errors resulting in incomplete or no UNL graphs. As the UNL generation system becomes more robust, the performance is expected to improve quickly. The overall result over all the datasets is given in the following table.

ALL	ALLnrm	Mean
0.3431	0.6878	0.3481

6 Conclusion and Future Scope

The UNL graph matching approach works well with grammatically correct sentences. The approach depends on the accuracy of the UNL generation system itself. With the increase in the robustness of the UNL generation system, this approach seems natural. Since, the approach is unsupervised, it does not require any training data. The matching algorithm can be extended to include the synonyms of the Universal Words while matching relations.

References

Mohanty, R. and Limaye, S. and Prasad, M.K. and Bhattacharyya, P. 2008. *Semantic Graph from English Sentences*, Proceedings of ICON-2008: 6th International Conference on Natural Language Processing Macmillan Publishers, India

UNL Center of UNDL Foundation 2005 Universal Networking Language (UNL) Specifications Version 2005 *Online URL:* http://www.undl.org/unlsys/unl/unl2005/

UNL enconversion system. 2012. Online URL: http://www.cfilt.iitb.ac.in/UNL_enco