Semantic Similarity Assessment using Universal Networking Language

A.Chitra, Anupriya Rajkumar
{*ctr.psg@gmail.com, anupriya_rajkumar@yahoo.co.in}

Professor and Head, CA Department, PSG College of Technology, Coimbatore- 641004Tamil Nadu, India [1] Professor,CSE Department, Dr.Mahalingam College of Engg. & Technology, Pollachi,- 642003 Tamil Nadu, India [2]

Abstract - Semantic similarity assessment is a key problem in Natural Language Understanding which finds wide application in Information Retrieval and Extraction. Determining whether two natural language text units are semantically equivalent is a challenging task. In this work, a machine learning approach based on matching of Universal Networking Language (UNL) forms has been proposed for semantic similarity assessment. Features which measure the relatedness of the UNL forms are used as input to a Support Vector Machine classifier to determine semantic equivalence. The performance of the system has been evaluated on the Microsoft Research Paraphrase Corpus with an accuracy of 71%. The suitability of the UNL matching scheme for handling multi-lingual inputs has also been demonstrated.

Keywords: Paraphrase Recognition, Cross Language Similarity, Support Vector Machine

1 Introduction

Human communication typically occurs through a multitude of natural language forms all of which are characterized by rich semantic variability and ambiguity. The establishment of Semantic Similarity between text units is a pivotal task in several applications such as Information Extraction, Question Answering and Summarization. Paraphrases and Entailment are two common forms of semantic similarity. Two text units are said to paraphrase each other, when exact semantic equivalence can be established between them. In text entailment, one of the inputs, usually the shorter one also termed as the hypothesis may be inferred from the longer unit or text .A logical solution for establishing semantic similarity would be to translate the input text units to an intermediate representation or Interlingua and then compare these. Such approaches typically rely on Symbolic Meaning Representations [1]. Graphs are constructed from the input text with words as nodes and edges representing the semantic relations between them. Similarity measures are then computed on these representations using additional resources such as Word Net. Universal Networking Language (UNL) is one such Interlingua. The UNL Project was launched by the Institute of Advanced Studies of the United Nations University in 1996 with the objective of eliminating language barriers [2]. The UNL project proposed an Interlingua based text representation to

ICCAP 2021, December 07-08, Chennai, India
Copyright © 2021 EAI
DOI 10.4108/eai.7-12-2021.2314540

support language independent computing applications and Internet interfaces. UNL represents sentences using hyper graphs, with nodes representing concepts and arcs the relationship between concepts. This paper describes an approach for determining whether two input sentences are semantically similar based on Universal Networking Language (UNL) representation. The advantage of such a system is that it can be used to establish cross-language similarity. The proposed system has been applied in accessing a cross-language FAQ for assessing the similarity between the question posed by the user and the queries available in the FAQ.

The key concepts involved in the current work are as follows: Paraphrase Recognition is the process of detecting whether a pair of text units is semantically equivalent and Cross-language similarity assessment handles inputs from different languages. The translation of natural language input to UNL form is termed as En conversion while the vice-versa is referred to as De conversion. UNL matching is the process of comparing the UNL forms of two sentences to determine the set of overlapping entities. Co-references in a text are multiple expressions which refer to the same entity and the process of identifying these expressions is termed as Co-reference Resolution. Word Sense Disambiguation deals with the determination of the specific sense of a word based on its context and Named Entity Recognition identifies the expressions in the input text which correspond to entities such as Person, Organization, Place, Time etc. Section 2 of the paper discusses about UNL and related work. Section 3 describes the proposed method of semantic similarity assessment and the results obtained on a standard Paraphrase corpus. Section 4 presents the system’s application and Section 5 concludes with future directions.

2 Literature Review

Paraphrase Recognition systems employ different techniques such as vector space models, surface string similarity, syntactic similarity, decoding and logic based approaches to establish semantic equivalence [1]. The usage of intermediate representations such as FrameNet’s frames or semantic roles from Prop Bank is one such technique. The semantic representations of the sentences are then compared to detect paraphrases. Burckhardt et al have carried out frame semantic analysis to represent the predicates and arguments in the sentence as frames and roles [4]. This process helps to overcome word-level variations of a semantic concept. Graph matching is then carried out by extracting various features from these semantic representations. Amoia et al [5] have extended the Xerox Incremental Parser with information from Verb Net and Word Net so as to produce the same semantic representation for paraphrases. Matching of a modified version of Conceptual Graphs which consists of concepts and relations between the concepts has been used by Boonthum et al for Paraphrase Recognition [6].Universal Networking Language has been proposed by United Nations University with the objective of developing universally usable computer interfaces. UNL represents the meaning of a sentence in the form of a semantic network with hyper-nodes. In the UNL semantic network, nodes represent concepts, and arcs represent relations between concepts [2]. The three basic components in the UNL representation are: Universal Words, Relations and Attributes. Universal words are English words used to represent simple or compound concepts. Relations indicate the relationship between two universal words while attributes provide additional information about the Universal words. The UNL representation of the sentence “This computer translates from English to Hindi” is as follows [7]:

agt(translate(icl>do).@entry, computer(icl>machine))
mod(computer(icl>machine), this)

src(translate(icl>do).@entry, english(icl>language))
gol(translate(icl>do).@entry, hindi(icl>language))

The process of converting a sentence given in natural language into UNL representation is termed as en conversion while the reverse process is termed as de conversion. Online servers which perform these tasks are available. One such server is the UNL Explorer which offers multi-lingual search, dictionary, UNL Ontology and UNL Talk services apart from conversion [8]. Other popular servers are the Russian-UNL server and the IITB-CFILT UNL server.

Sentences in any language have common UNL representation. Hence UNL can be used as an intermediate language for finding semantic similarity between the sentences. Pakray et al (2010) have developed a UNL based text entailment recognition system [9]. Similarity assessment is done by assigning scores depending on the extent of relation matches. Two relations are said to match exactly if the entire relation with all its arguments match. Wordnet synonyms and expanded relations are used for identifying approximate matches. The system has been evaluated using the RTE-3(Recognizing Textual Entailment) and RTE-4 datasets with precision and recall of 60%. In [10] Pakray et al have extended the UNL matching system for the Answer Validation task by grouping similar relations and considering word synsets as well as Named Entity matches.

Singh et al [11] have assessed sentence similarity by performing UNL matching. A three stage scoring system has been used where attribute and word matching scores contribute to universal word scores which in turn are used for calculating relation scores. F1-score is calculated from Precision and recall scores which are computed by dividing the aggregate relation score by the number of relations in the UNL forms of each of the input sentences. The system has achieved an accuracy of only 19.36% on the Microsoft Research Paraphrase Corpus (MSRPC) due to difficulties in UNL conversion. Dan et al [12] have used lexical and syntactic features in addition to semantic features extracted from UNL graphs for measuring similarity. A linear regression model was built from training data and then used for prediction on the test data set with good results only for short sentences. Goel and Kumar [13] have proposed various NLP applications for the UNL representation including Semantic Textual Similarity, Machine Translation, Text Summarization, Clustering, Sentiment Analysis etc. Several Indian language enconverters and deconverters have been developed for languages such as Bengali, Tamil and Punjabi. Sitender and Bawa [14] have developed a Sanskrit to UNL enconverter system which has recorded a good BLEU score.

From a survey of related work it can be observed that though UNL is a universal intermediate representation, relatively less work has been carried out with respect to text similarity assessment using UNL. Further there is scope for improvement in the performance of such UNL matching systems available currently.

3 UNL Based Paraphrase Recognition

In this work an alternate scheme which employs a machine learning classifier has been proposed for UNL matching. Various features extracted from the UNL forms of input sentences are used to classify whether the two text units are semantically similar. Word Sense Disambiguation and Co-reference Resolution have been included as additional options and their effect on UNL Matching has also been investigated. Fig I shows the stages of the proposed UNL matching system.

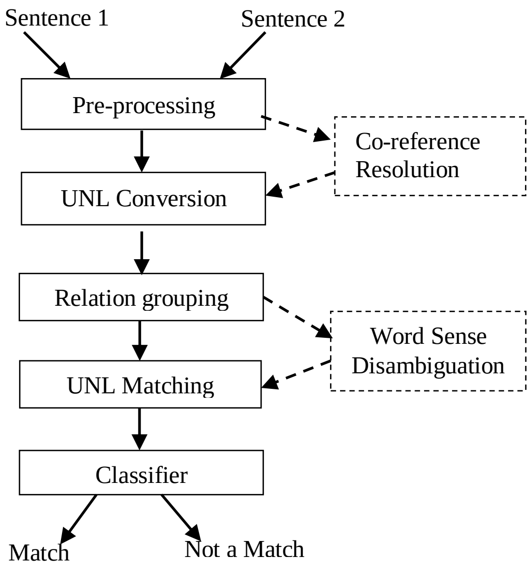

Fig. 1Stages of UNL based Paraphrase Recognizer

The input sentences are subjected to pre-processing to eliminate punctuations and word contractions. Co-reference resolution is applied to replace pronouns in the input sentence with their corresponding nouns. This step has been applied before UNL conversion to ensure that co-referring nouns and pronouns within a sentence map to the same Universal words. The sentences are then converted to UNL form. To facilitate the matching of UNL forms, the UNL relations which have distinct labels but are similar in terms of the underlying concept are grouped together. Further, for the Universal words corresponding to nouns and verbs, Word Sense Disambiguation is applied to determine the specific sense in which a word is used. As this step requires the input sentence to be matched against the various senses of a word extracted from Word Net, it would not be feasible for cross-language inputs which do not have Word net like dictionaries. The UNL forms of the input sentences are then compared to extract various features which reflect the degree of similarity between the UNL forms. Since assessing semantic equivalence by using simple rules and thresholds is tedious as well as inaccurate, a machine learning classifier has been used. The extracted features are used by the classifier to detect whether the input sentences are semantically similar.

3.1 Pre-processing and Co-reference Resolution

During the Pre-processing stage, punctuations are removed and contractions such as “aren’t”, “didn’t” are expanded. This is followed by Co-reference resolution which is the process of identifying all expressions that refer to the same entity. For example in the sentence “Ram went to the shop and he bought a book”, the pronoun ‘he’ refers to ‘Ram’. The Stanford Co-reference Resolution System [15] has been used to resolve co-references in the input sentences.

3.2 UNL Conversion and Processing

UNL En-conversion has been carried out using the Russian UNL Converter. The Russian UNL convertor is an online language server which converts English and Russian language sentences into UNL form. Due to its better availability and the ability to handle longer sentences [16], the Russian Converter has been used in this work. Relations in UNL are also referred as links and indicate the relationship between the universal words. There are 39 distinct relation labels in UNL, with some of them being related. For instance agent(‘agt’) refers to the initiator of an action on whom the sentence focuses whereas partner(‘ptn’) indicates a non-focussed initiator [2]. In this work, related relations have been grouped together to facilitate better matching similar to [11]. Sample relation groups have been shown in Table I. All relations within the same group are said to match approximately.

TABLE I
SAMPLE RELATION GROUPS

Time-related	Place-related	State
Tim, tif, tmt	plc, plf, plt	Src, gol, via

3.3 Word Sense Disambiguation and Similarity Calculation

Word Sense Disambiguation (WSD) is the process of determining the exact sense of a word based on its context. Though the Constraint list attached to a Universal Word aims to restrict the sense of the word, the same word may have different meanings in different contexts. The adapted Lesk algorithm [17] has been used here. The Lesk algorithm determines all possible senses of the target word as well as their glosses from WordNet. A target word may have multiple senses, each of which has a corresponding gloss or textual definition. The sense whose gloss has the highest degree of word overlap with the context is chosen as the best sense of the word. Here the context refers to the sentence containing the target word. In the current work, the effect of WSD on UNL matching has also been studied.

In the cases where the Universal Words do not match exactly, similarity between the words is computed using WordNet. The Jiang-Conrath measure which assesses the similarity between two words in terms of the information content of the given words and their lowest common subsumer in the WordNet hierarchy has been used [18]. The Jiang-Conrath score ranges between 0 and 1, with the value being closer to 1 as the similarity increases. Word pairs with a score greater than 0.15 have been taken to be similar words.

3.4 UNL Matching

In order to assess semantic similarity, various features are computed from the UNL forms of both the sentences by matching the Relations, Universal Words and UNL attributes. The features and the rules for their computation are listed in Table II where S1, S2 are the input sentences, R1, R2 represent relations and UW1-UW4 are the four universal words from R1, R2. Additionally the Named Entities in both sentences have been identified using the Stanford Named Entity Recognizer [19] and compared to generate a pair of features.

TABLEII
UNL Matching Features

Feature	Rules for Computation
Simple_Rel_Precision	Number of common relations between S1 and S2 divided by the number of relations in S1, S2
Simple_Rel_Recall	
UW_Precision	Number of matching Universal word pairs between S1 & S2 divided by number of Universal word pairs in S1, S2
UW_Recall	

Overall_Rel_Precision	If R1=R2 or if both are in the same group, the pairs UW1, UW3 and UW2, UW4 are compared. If they match exactly or if they are similar R1 and R2 are said to match. Number of matching relations is divided by number of relations in S1, S2.
Overall_Rel_Recall	
Named_Entity_Precision and Recall	Number of matching named entities divided by number of named entities in S1, S2

All of the above features are applicable for cross-language inputs. WSD has been provided as an additional option which can be disregarded in the case of cross-language inputs.During Universal word matching with the Word Sense Disambiguation option, when two words do not match exactly, the following steps are performed:

- Specific sense of each Universal word is determined by applying Word Sense Disambiguation
- Synsets corresponding to the specific senses are obtained
- If the synsets overlap then the Universal words are said to match.

A Support Vector Machine (SVM) Classifier has been used to classify the sentences as positive or negative cases of paraphrases using the features extracted from the input sentence pair. Support Vector Machine is a supervised learning algorithm which classifies data by constructing a Maximal Margin Hyperplane [20]. SVMs handle linearly inseparable data by mapping them to a higher dimensional space by using kernel functions. In this work, the LibSVM tool [21] has been employed to implement a nu-Classifier with a radial basis function kernel.

3.5 Results

The Microsoft Research Paraphrase Corpus (MSRPC) consisting of 5801 pairs of sentences has been used for evaluating the performance of the UNL matching system. It consists of 3900 positive cases of paraphrases and 1901 negative cases [22]. The corpus is divided into a training set with 4076 sentence pairs and test set with 1725 pairs. The performance of the system has been assessed using the Microsoft Research Paraphrase Corpus, in terms of Accuracy and F-measure.The performance of the proposed UNL matching scheme has been compared with that of Singh et al's system [11] which has recorded an accuracy of 19.36% when the IITB-CFILT UNL converter has been used. For benchmarking the proposed system, the existing system performance has been reassessed by using the Russian UNL converter. One of the difficulties faced in the existing system is that of fixing the threshold suitably. Variations of the existing system which were tried include: using the originally prescribed threshold of 0.5 and classification using a Support Vector Machine Classifier which avoids the need for using a threshold. The obtained accuracy values have been listed in Table III.

TABLE III
Performance Evaluation of UNL Matching system

System Variant		Accuracy %	F-measure %
Existing System	Threshold = 0.5	59.65	66.63
	SVM	66.49	78.02

	Classification		
Proposed System	All features	71.01	80.95
	Without Named Entity features	70.55	80.89
	After Co-reference resolution	66.49	78.46
	Word Sense Disambiguation	70.78	80.91

With respect to the proposed system, experiments were conducted by considering various combinations of the features listed in Table II and options described in earlier sections such as Co-reference resolution, Word Sense Disambiguation. Of the features listed in Table II, the best individual performance of 70.84% Accuracy was registered by Universal Word features. This can be attributed to the fact that similar sentences or paraphrase pairs in MSRPC exhibit a considerable degree of word overlap. This was followed by the Overall Relation Precision and Recall and finally Simple Relation features with accuracies of 68.23% and 67.53% respectively. This difference is due to the fact that the overall relation features permit the matching of relations within the same group and do not require the candidate relations to be exactly the same. Since Named Entity features do not include any of the UNL entities such as UWs or relations, they have not been used as stand-alone inputs.

From the results of the experiments it can be observed that the proposed UNL matching system which uses all the eight features has the best overall performance. An increase in accuracy of more than 11% is observed when compared to the existing approach of [11]. A notable aspect is that combining the scoring mechanism of the existing system with a machine learning approach serves to improve the accuracy considerably.

Experiments were also conducted by performing Co-reference Resolution and WSD independently. Other aspects that can be inferred from the results are that:

- the usage of Named Entity features improves the performance of the system as paraphrases tend to share more Named entities
- When Co-reference resolution was carried out on pronouns before computing the features listed in Table II a drop in performance was observed. This can be attributed to the fact that in some cases the references are wrongly resolved as shown below:

Original Sentence: But under cross-examination by O'Donnell's attorney, Lorna Schofield, Toepfer conceded she had ignored many of O'Donnell's suggestions and projects.

After Co-reference resolution: But under cross-examination by O'Donnell's attorney, Lorna Schofield, Toepfer conceded O'Donnell's had ignored many of O'Donnell's suggestions and projects.

Additionally with respect to semantic similarity assessment, co-references across sentences rather than within the same sentence is more of an issue as can be seen from the example given below:

Sentence 1: The two had argued that only a new board would have had the credibility to restore El Paso to health.

Sentence 2: He and Zilkha believed that only a new board would have had the credibility to restore El Paso to health.

- Applying Word Sense Disambiguation on universal words (specifically nouns and verbs) during UNL matching, leads to a slight drop in accuracy due to the reason that considering specific senses of words is more restrictive.

The inclusion of word overlap features resulted in an increased accuracy of 73%. But this requires both the input sentences to belong to the same language. Since the objective here is to develop a system for measuring similarity between inputs from different languages, such features have been disregarded. However, the best performing systems for Paraphrase Recognition have registered an accuracy of more than 75% on the MSRPC [1]. A possible direction for future improvement is the identification of additional features from the UNL forms.

4 Cross-language FAQ Access

Semantic similarity assessment has wide impact in many real time applications such as Information retrieval, Summarization and Automatic Question Answering. The applicability of the proposed UNL matching system for accessing a cross-language Frequently Asked Questions (FAQ) database has been investigated. Access to Russian language FAQ has been demonstrated since the Russian UNL Converter has been used. Though multi-lingual search and translation services are supported in the UNL Explorer interface [8], a notable aspect of the current interface is that it supports sentence level queries and computes the semantic similarity between the UNL forms of the user query and the queries available in the database.

The FAQ of Ozon an online megastore which consists of 104 questions addressing various queries pertaining to online shopping has been considered. The user’s English language query is converted to UNL form and compared with the UNL version of the 104 queries in the FAQ. The Russian answer for the matching query is fetched and then translated to English using Google Translation services and presented to the user. The advantages of this system are that it permits access to cross-language FAQs and that the query need not be an exact translation of the original query as the system checks for paraphrases and not exact matches.

Conclusion

Effective mechanisms are required for Semantic Similarity Assessment which is a challenging task. The original contribution of this paper is the design of a machine learning approach which operates on features extracted from the UNL forms of input sentences for Semantic Similarity assessment. Evaluation experiments prove that the system performs better than the existing UNL matching approach. An important aspect of the system is its suitability for measuring similarity between sentences of different languages because UNL has been used as an intermediate representation. This aspect has been demonstrated by deploying the system for accessing cross-language FAQs. Future work includes the incorporation of additional features extracted from the UNL forms to improve the accuracy.

References

[1] Androutsopoulos I and Malakasiotis P A Survey of Paraphrasing and Textual Entailment Methods. Journal of Artificial Intelligence Research, vol .38, pp:135-187. 2010

[2] UNL Center, UNDL Foundation The Universal Networking Language (UNL) Specifications. Version 3,Edition 2. 2003

[3] Cardeñosa J, Gallardo C, and Iraol L Using an Interlingua for Document Knowledge Representation.In Proc. of 4th EUSFLAT Conference, Spain, pp. 1231-1236. 2005

[4] Burchardt A, Reiter N, Thate S, and Frank A A Semantic Approach to Textual Entailment: System evaluation and task analysis.In Proc. of the ACL-PASCAL

Workshop on Textual Entailment and Paraphrasing Association for Computational Linguistics, Prague, pp. 10-15. 2007

[5] Amoia M and Gardent C Recognition of alternation paraphrases: a robust and exhaustive symbolic approach.In Proc.of Knowledge and Reasoning for Answering Questions-KRAQ'05, Edinburgh, pp.57-60. 2005

[6] Boonthum C,Toida S and Levinstein I Paraphrasing Recognition through Conceptual Graphs. Technical Report, Computer Science Department, Old Dominion University, Norfolk.2003

[7] Grishkyan Y Upon Comparison of Some Online UNL Modules. In Proc. of UNL Workshop of CSIT 7th International Conference, Yerevan, pp.293-296. 2008

[8] Uchida H, Zhu M, Salam KMA UNL Explorer.In Proc. of COLING 2012, Mumbai, pp.453-458. (2012)

[9] Pakray P, Poria S, Bandyopadhyay S and Gelbukh A Semantic Textual Entailment Recognition using UNL. Polibits vol.43, pp. 23-27.2011

[10] Pakray P Answer Validation through Textual Entailment. Lecture Notes in Computer Science N 6609, Springer, pp. 359-364. 2011

[11] Singh J, Bhattacharya A and Bhattacharyya P Semantic Textual Similarity using Universal Networking Language graph matching.In Proc. of the Sixth International Workshop on Semantic Evaluation SemEval’12, Association for Computational Linguistics, Montréal, pp. 662-666. 2012

[12] Dan A and Bhattacharyya P CFILT-CORE: Semantic Textual Similarity using Universal Networking. In Proc. of the Second Joint Conference on Lexical and Computational Semantics, vol. 1, Atlanta, pp. 216–220.2013

[13] Kumar P and Goel K Universal networking language: A framework for emerging NLP applications.In 1st India International Conference on Information Processing (IICIP), pp. 1-6. 2016

[14] Sitender and Bawa S SANSUNL: A Sanskrit to UNL Enconverter System. IETE Journal of Research, vol. 67, pp. 117-128. 2021

[15] Lee H, Peirsman Y, Chang A, Chambers N, Surdeanu M and Jurafsky D Stanford's multi-pass sieve co-reference resolution system at the CoNLL-2011 shared task. In Proc. of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, Association for Computational Linguistics, USA, pp. 28-34. 2011

[16] UNL Russian Module Server, www.unl.ru, www.proling.iitp.ru/deco

[17] Navigli R Word Sense Disambiguation: A Survey. ACM Computing Surveys, vol. 41(2), pp. 1-69.2009

[18] Jiang J and Conrath D W Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy.In Proc. of International Conference on Research in Computational Linguistics, Taiwan, pp.19-33. 1997

[19] Finkel R, Grenager T and Manning C Incorporating non-local information into information extraction systems by Gibbs sampling. In Proc. of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 363-370.(2005)

[20] Shawe-Taylor J and Cristianini N Support Vector Machines and other kernel-based learning methods. Cambridge University Press 2000

[21] Chang C C and Lin C J LIBSVM: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology, vol.2, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. 2011

[22] Dolan B, Quirk C and Brockett C Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources.In Proceedings of 20th Int. Conf. on Comp. Linguistics, Geneva, pp.350–356. 2004