Indonesian Journal of Electrical Engineering and Computer Science
Vol. 1, No. 2, February 2016, pp. 371 ~ 374
DOI: 10.11591/ijeecs.v1.i2.pp371-374

Semantic Similarity/Relatedness for Cross Language Plagiarism Detection

Hanane Ezzikouri*, Mohammed Erritali, Mohamed Oukessou
LMACS laboratory, Mathematics Department,
Faculty of sciences and techniques Sultan Moulay Slimane University Beni-Mellal, Morocco
*Corresponding author, e-mail: ezzikourihanane@gmail.com

Abstract

Generally utterances in natural language are highly ambiguous, and a unique interpretation can usually be determined only by taking into account the context in the utterance occurred. Automatically determining the correct sense of a polysemous word is a complicated problem especially in multilingual corpuses. This paper presents an application programming interface for several Semantic Relatedness/Similarity metrics measuring semantic similarity/distance between multilingual words and concepts, in order to use it after for sentences and paragraphs in Cross Language Plagiarism Detection (CLPD); using WordNet for the English-French and English-Arabic multilingual plagiarism cases.

Keywords*: Semantic similarity, CLPD, Plagiarism.*

1. Introduction

Plagiarism can be defined as the reuse of someone else's ideas, results, or words without acknowledging the original source. Cross-Language Plagiarism Detection (CLPD) consists in discriminating semantically similar texts independent of the languages they are written in, when no reference to the original source is given. CLPD case takes place when we deal with unacknowledged reuse of a text involving its translation from one language to another [1].

CLPD issue has acquired pronounced importance lately since semantic contents of a document can be easily and discreetly plagiarized through the use of translation (human or machine-based).

Arabic is a Central Semitic language, it belongs to the Afro Asiatic family, Arabic has much specificity which makes it very different from other Indo-European languages. Detecting plagiarism in Arabic documents is particularly a challenging task and it becomes even harder, as translation is often a fuzzy process that is hard to search for, because of the complex linguistic structure of Arabic. In spite of the fact that many researches were conducted on plagiarism detection in the last decades, those concerning the Arabic language text remain quite limited and addressed especially to monolingual plagiarism.

Similarity is a fundamental and widely used concept. An important number of similarity measures have been proposed in the last few years.

The similarity between two subjects (e.g.: A and B) is related to their commonality/differences. The more commonality/ differences they share/have, the more/less similar they are. While Semantic Similarity Semantic similarity (Ss) [2] refers to similarity between two concepts in a taxonomy such as the WordNet, where the idea of Semantic similarity between them is based on the likeness of their meaning or semantic content as opposed to similarity which can be estimated regarding their syntactical representation [3]. Semantic similarity is often confused with semantic relatedness, where the second one includes any relation between two terms. For example, "car" and "bus" are similar in that they are connected via a relation with "vehicle", but is only related to "road" and "driving".

Natural language utterances are, in general, highly ambiguous, because of the multiple possible meaning or senses that words may have (polysemous) or malapropism which is the confounding of an intended word with another word of similar sound or similar spelling that has a quite different and malapropos meaning, and interpretation can generally be determined only

Received October 12, 2015; Revised January 17, 2016; Accepted January 30, 2016

by taking into account the context in which the utterance occurred. However, algorithms/programs do not have the benefit of human's vast experience of the language.

The steps of Cross-Language Plagiarism Detection process were defined by [4], authors put up some strategies of heuristic retrieval and evaluate the performance of the models for the detailed analysis. Although many studies were directed on plagiarism detection in the last years, those concerning the Arabic language text remain quite limited. Works in this area are those of Alzahrani et al. [5], Menai et al. [6] and others [7, 8]. All of them addressed the monolingual external approach.

N. Abdul Jaleel, et al. [9] works on statical transliteration for English-Arabic Cross Language Information retrieval (CLIR), authors worked with n-gram model and evaluate the statistically-trained model and a simpler hand-crafted model on a test set of named entities from the Arabic AFP-Corpus and demonstrate that they perform better than online translation sources.

Imene Bensalem and al. [10] Work on Arabic intrinsic plagiarism detection. They presented a set of preliminary experiments on intrinsic plagiarism detection in Arabic text using Stylysis tool and a small corpus. Their approach consists in testing whether some language-independent stylistic features are effective or not to discriminate between plagiarized and not plagiarized sentences, the results they found is that average word length and average sentence length are not reliable stylistic discriminator of Arabic text.

2. Used Metrics

2.1. HSO [11]

A malapropism can be defined as a correctly spelled word unsuitable with the context where it is used, because it is may a spelling error of another intended word. The detection of malapropisms relied, on a big quantity, on lexical chains presenting words of a semantic continuity.

Hirst&Onge fixed a mechanism that generates spelling replacements that can be used to generate replacement candidates for a malapropism, basing their arguments on the fact that words that cannot be used with other words can be considered as potential malapropisms. The proposed algorithm uses the WordNet thesaurus to automatically quantify semantic relations between words, in WordNet, a word may have one to many synset, each corresponding to a different meaning. When we look for a relation between two different words, we consider the synsets of all the senses of each word, looking for a possible connection between some meanings of the two word.

2.2. Lesk [12]

Word sense disambiguation is the task of identifying the intended meaning of a given target word from the context in which it is used. In WordNet each concept (or word sense) is defined by a short gloss. A super-gloss of a concept is an expanded gloss that concatenate other glosses that are connected to it via some WordNet relation.

The Adapted Lesk measure was developed to overcome the problem of short definitions in most dictionary, which was an interest to (Lesk, 1986) when he present the notion of involving definition overlaps for word sense disambiguation. In the A-Lesk measure, similarity between two word senses (concepts) attributed by finding and scoring intersections between the glosses of two concepts. The bigger number of intersection gloss word is, indicates a stronger relation, the bigger similarity value between two concepts.

2.3. LCH [13]

The LCH similarity/relatedness measure (Leacock and Chodorow) is:

$$SIM_{LCH} = -\log\left(\frac{length}{2*D}\right)$$

Where:

Length is the length of the shortest path between the two synsets (using node-counting)
D is the maximum depth of the taxonomy

LCH measure is very sensitive to the presence or absence of a unique root node, is very sensitive to the presence or absence of a unique root node because it consider the depth of the taxonomy in which the synsets are found.

2.4. LIN [3]

The LIN similarity measure is:

$$vSIM_{LIN} = 2 \times \frac{IC(LCS)}{IC(Concept1)+IC(Concept2)}$$

Where IC(x) is the information content of x, And LIN similarity verify $0 \leq SIM_{LIN} \leq 1$

If there is any lack of data or the information content of any of either concept1 or concept2 is0, then 0 is returned as the similarity score.

2.5. WUP [14]

The WUP similarity/relatedness measure (Wu & Palmer) is:

$$vScore = 2 \times \frac{depth(LCS)}{depth(S1)+depth(S2)}$$

Where the depths of the two synsets in the WordNet taxonomies, along with the depth of the LCS (Least Common Subsumer (LCS).

3. Application

The Semantic Relatedness/Similarity calculus process is divided into a number of smaller sub-tasks, each of which is using metrics of relatedness/similarity. Each of the sequential sub-tasks or steps accepts data from a previous stage, performs a transformation on the data, and then passes on the processed data structures to the next step. In the development of our system, we did use some Semantic Relatedness/Similarity algorithms that exist the java API WordNet Similarity for Java (WS4J) with some modification to makes them suitable for Arabic.

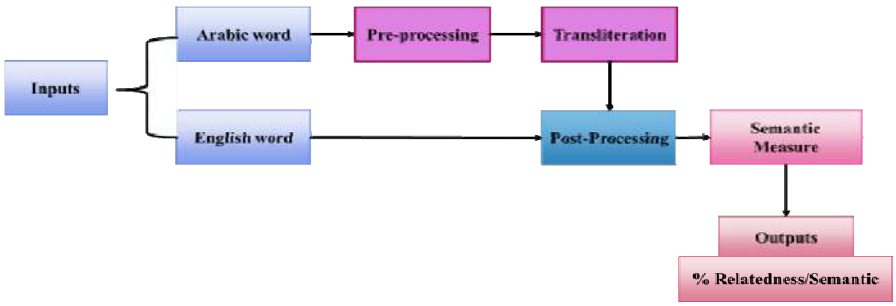

Figure 1. A generalized process for Arabic/English semantic similarity measure

3.1. Pre-processing

Some optional pre-processing should be performed on the data structures entered by the user. This would include tasks such as dealing with “Harkat/Tashkeel” from the user input, which is the process of removing it from the Arabic concepts, based on heuristics and some algorithms.

3.2. Transliteration

We have used the Java port of the homonym product developed in Perl by Tim Buckwalter, it works with a transliteration of the Arabic word. This transliteration uses

Buckwalter's transliteration system. It includes Java classes for the morphological analysis of Arabic text files, whatever their encoding.

3.3. Proposed System

We have developed a graphical interface to conveniently access the system. The GUI is written in java. The interface allows the user to input words, and to submit for semantic similarity calculation.

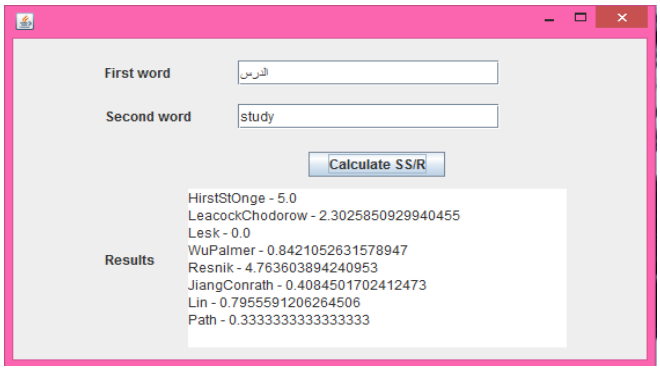

Figure 2. Proposed System

4. Conclusion

As part of this work, we developed a system to calculate semantic similarity/relatedness in a Arabic-English. Our objectives are to to extend the system to englobe scientific articles and academic researches. We believe that the validation of the results requires further experiments in time.

References

[1] Potthast M, Barrón-Cedeño A, Stein B, Rosso P. Cross-language plagiarism detection. *Language Resources and Evaluation*. 2011; 45(1): 45-62.
[2] Philip Rensik. *Using information content to evaluate semantic similarity in a taxonomy*. In Proceedings of IJCAI-95. Montreal, Canada. 448-453.
[3] Dekang Lin. An Information-Theoretic Definition of Similarity. *ICML*. 1998: 296-304.
[4] Pereira RC, et al. *A new approach for cross-language plagiarism analysis*. In Multilingual and Multimodal Information Access Evaluation. Springer Berlin Heidelberg. 2010: 15-26.
[5] Alzahrani, Salha, Naomie Salim. Fuzzy semantic-based string similarity for extrinsic plagiarism detection. Braschler and Harman. 2010.
[6] Menai MEB. Detection of plagiarism in Arabic documents. *International journal of information technology and computer science (IJITCS)*. 2012; 4(10): 80.
[7] Zitouni, Abdelaziz, et al. Corpus-based Arabic stemming using N-grams. *Information Retrieval Technology*. Springer Berlin Heidelberg. 2010: 280-289.
[8] Siddiqui, Muazzam Ahmed, et al. Developing An Arabic Plagiarism Detection Corpus. Grant No. 11-INF-1520-03.
[9] Abdul Jaleel, Nasreen, Leah S Larkey. *Statistical transliteration for English-Arabic cross language information retrieval*. Proceedings of the twelfth international conference on Information and knowledge management. ACM. 2003.
[10] Bensalem, Imene, Paolo Rosso, Salim Chikhi. *Intrinsic plagiarism detection in Arabic text: Preliminary experiments*. II Spanish Conference on Information Retrieval (CERI'12). 2012.
[11] Graeme Hirst, David St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. WordNet, edited by Christiane Fellbaum, Cambridge, MA: The MIT Press. 1995.
[12] Satanjeev Banerjee, Ted Pederson. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. 2002.
[13] Claudia Leacock, Martin Chodorow. Combining Local Context and WordNet Similarity for Word Sense Identification. WordNet: An Electronic Lexical Database, Publisher: MIT Press. 265-283.
[14] Wu Z, Palmer M. *Verbs semantics and lexical selection.* Proceedings of the 32nd annual meeting on Association for Computational Linguistics Association for Computational Linguistics. 1994: 133-138.