Using WordNet-based semantic similarity measurement in External Plagiarism Detection

Notebook for PAN at CLEF 2011

Yurii Palkovskii, Alexei Belov, Iryna Muzyka

Zhytomyr State University, SkyLine. Inc, "MARS" p.e.
palkovskiy@yandex.ru

1 Introduction

Continuing our previous work started at PAN 2009 and PAN 2010 [7] we considered further research options based on the achieved baseline of the best performing algorithms. The research done by Potthast et al. [4] presented a sliced view of the presented approaches showing their performance on specific corpus metrics - external\intrinsic, obfuscation strategies (none, artificial high\low, simulated, translated), topic match, case length and document length thus defining the baseline for further studies. A brief analysis of the above named results [1,3] shows that there exists a direct correlation between the obfuscation degree (method) and the achieved performance.

Table 1. Detection Efficiency (DE% Recall::Precision) in relation to "no obf." as a base value:

obfuscation type:	1 result			2 result		
	P	R	**DE%**	P	R	**DE%**
no obfuscation	94	96	100%	78	86	100%
low	93	92	98%	81	85	103%
high	93	75	98%	76	76	97%
translation	92	70	97%	58	47	74%
simulated	33	18	35%	19	22	24%

I the above table the complexity of obfuscation rises from "no obfuscation" to "simulated" type. An important note should be made - we rearranged "translated" plagiarism with "simulated" treating the latter as the most advanced form of obfuscation due to the fact that the exact "translation" mechanism used for cases generation, followed the "bag of words" pattern, instead of Mechanical Turk human translation. The achieved score proves the above idea and opens a new vector for further developing the corpus including human made translations.

Additionally, some minor criticism has been put forward in relation to the PAN plagiarism detection "definition of plagiarism encompasses much more than just character sequence matching" [2]. This particular opinion and the performance baseline difference mentioned above led us to the idea of switching from the quantum based statistical n-gram probing (n-gram fingerprinting and TFIDF measurement) to semantic similarity measurement (SSM). Our primary hypothesis stated that SSM used as a main comparer must be more resilient to simulated plagiarism and translated plagiarism as it relies on sense quantum instead of character quantum. The success of the SSM heavily depends on the exact method of plagiarism construction. Artificial plagiarism, starting from low to high obfuscation levels and including translated plagiarism, constructed by the PAN Random Plagiarist follows "bag-of-words" approach [4] and lacks many features that differentiate it from "human generated rewrite" thus giving the ground for the above mentioned criticism. Among these missing features we can name correct word order, correct grammar, writing style, word choice, etc. and more importantly - the sense structure behind the text that conveys sense itself. In an effort to overcome these limitations PAN 2010 has launched "simulated" plagiarism sections into the corpus. Generated by Mechanical Turk with multiple quality validation followed, these cases have become our primary target for SSM experiments as they represented the real world plagiarism cases generated by human, that convey all the previously missing features mentioned above.

2 External Plagiarism Detection

Due to serious time constraints we decided to use the previously developed system for candidate document retrieval and focus our research on the SSM comparer development. Selecting the SSM as our primary document similarity comparison we were aware that most probably we will not be able to achieve better or even comparable results in relation to the existing PAN baseline due the following factors - the majority of plagiarism cases are not simulated but artificial plagiarism, extremely heavy performance load on the comparer that resulted in excessive processing time requirements. Still we strongly believe that SSM is the future of plagiarism detection so we decided to pursue this particular method to discover its possible benefits.

Initially we planned to build our SSM comparer from scratch building it up around the idea of measuring the distance via WordNet synsets but when we discovered the Troy Simpson's project on SSM [5] we decided to use this project as a foundation of your own prototype. Our SSM text comparer comprises several open-source projects, namely:

- WordNet 2.1
- WordNet.Net an open-source .NET Framework library for WordNet
- C# Porter Stemmer implementation
- C# Brill Tagger implementation
- SSM Words Comparer by Troy Simpson and Thanh Dao
- SSM Sentence Comparer by Troy Simpson and Thanh Dao

The strategy to capture semantic similarity between two sentences. Given two sentences X and Y, we denote m to be length of X, n to be length of Y. The major steps can be described as follows:

1. Tokenization.

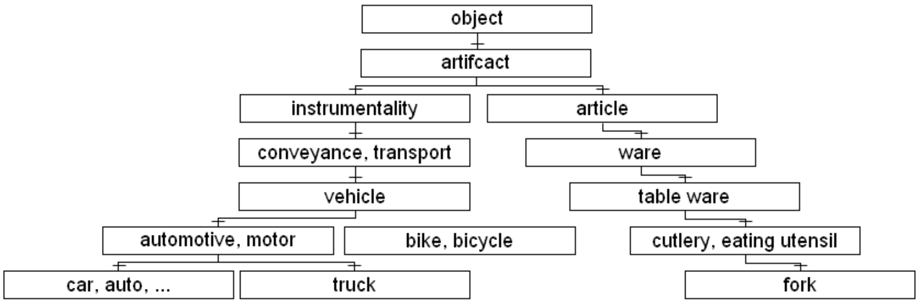
2. Perform word stemming.
3. Perform part of speech tagging.
4. Word sense disambiguation.
5. Building a semantic similarity relative matrix R[m, n] of each pair of word senses, where R[i, j] is the semantic similarity between the most appropriate sense of word at position i of X and the most appropriate sense of word at position j of Y. Thus, R[i,j] is also the weight of the edge connecting from i to j.
6. We formulate the problem of capturing semantic similarity between sentences as the problem of computing a maximum total matching weight of a bipartite graph, where X and Y are two sets of disjoint nodes

7. The match results from the previous step are combined into a single similarity value for two sentences. There following strategy was used:

$$\frac{2 \times Match(X, Y)}{|X| + |Y|}$$

According to Dhanh Tao approach [5,6], the path length-based similarity measurement example the length between car and auto is 1, car and truck is 3, car and bicycle is 4, car and fork is 12.

Figure 1. Hyponym taxonomy in WordNet used for path length similarity measurement [5,6]:



At text level sliding window approach is used to utilize the SSM comparer. This approach was inspired by the style changing function measurement approach used in PAN 2010 intrinsic plagiarism detectors [3].

3 Evaluation

For detailed evaluation of the progress we developed an application to prepare a pre-selected corpora from the test corpus of PAN 2010 using the meatacriteria of the corpus itself. As we focused on the simulated plagiarism mainly - we trained our SSM comparer on this specific corpus. Taking into the consideration large amount of processing to be done and the lack of code optimization we need to further investigate the differences between our intermediate results that were achieved on our sub-corpus and the test corpus of the 2011. We suspect that PAN 2011 corpus has some specifics that presupposed the baseline change from 79% to 50% that will be announced at the workshop.

4 Conclusion

Concluding our research, we would like to note that using SSM as plagiarism detection mechanism has to be much improved to achieve competitive results. Our research is a small step towards the better understanding of semantic similarity usage at large scale data processing. Migration to a more productive languages and or infrastructure with possible cauterization may yield better results to face the conditions of the PAN competition and we strongly believe that we will be able to launch full semantic search for the next PAN Competition.

Among further research vectors we can name the following:

1. Better WSD methods.
2. Overall speed\performance optimization.
3. Further research on the problem of Semantic Hashing and Search.
4. Further research on the problem of semantic normalization.
5. Better definition of meta parameters used via machine learning.

We suspect that SSM will do best on human translated plagiarism but this needs to be proved on a larger corpus yet.

Bibliography

[1] Cristian Grozea and Marius Popescu. Encoplot—Performance in the Second International Plagiarism Detection Challenge: Lab Report for PAN at CLEF 2010. In Braschler et al. ISBN 978-88-904810-0-0

[2] Debora Weber-Wulff, "Plagiarism Detection Competition" copy-shake-paste.blogspot.com. 2009. 21 June.2011.

[3] Markus Muhr, Roman Kern, Mario Zechner, and Michael Granitzer. External and Intrinsic Plagiarism Detection using a Cross-Lingual Retrieval and Segmentation System: Lab Report for PAN at CLEF 2010. In Braschler et al. [2]. ISBN 978-88-904810-0-0.

[4] Martin Potthast, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeño, and Paolo Rosso. Overview of the 1st International Competition on Plagiarism Detection. In Benno Stein, Paolo Rosso, Efstathios Stamatatos, Moshe Koppel, and Eneko Agirre, editors, SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09), pages 1–9. CEUR-WS.org, September 2009. URL http://ceur-ws.org/Vol-502.

[5] Thanh Dao. "An improvement on capturing similarity between strings" www.codeproject.com. 2005. 29 Jul. 2011. http://www.codeproject.com/KB/recipes/improvestringsimilarity.aspx

[6] Troy Simpson, Thanh Dao. "WordNet-based semantic similarity measurement" www.codeproject.com. 2005. 1 Oct. 2011. http://www.codeproject.com/KB/string/semanticsimilaritywordnet.aspx

[7] Yurii Palkovskii, Alexei Belov, and Irina Muzika. Exploring Fingerprinting as External Plagiarism Detection Method: Lab Report for PAN at CLEF 2010. In Braschler et al. [2]. ISBN 978-88-904810-0-0.