A new approach for searching translated plagiarism

Máté Pataki
Computer and Automation Research Institute, Hungarian Academy of Sciences
mate.pataki@sztaki.hu

Keywords: external plagiarism detection, translated plagiarism

1. Introduction

Detecting plagiarism and similarity between documents written in the same language can be done with high precision with today's top search systems; there are both free e.g. Plagiarisma (2012), Copyscape (2012) and commercial ones available to use e.g. PlagAware (2012), turnitin (2012). With the spread of foreign language knowledge and the growing number of international students, a new form of plagiarism, translated plagiarism gained ground. When a work or part of it is translated to another language without giving credit to the original author, it is called "translated" or "cross language" plagiarism.

In 2010 we started a one-year research project to be able to detect cross language plagiarism cases. As Dr. Debora Weber-Wulff (2010), professor at the HTW Berlin and the author of the *Copy, Shake, Paste* blog (2012), put it, at that time „the biggest gap in all the plagiarism checkers was the inability to locate translated plagiarism" (Bailey, 2011). Most current approaches use machine translation to detect similarity between texts written in different languages. At the last International Plagiarism Conference Storm said: "The first step in offering a translated plagiarism detection service is to find a partner that offers machine translation." (Storm, 2010), but that works only if a good quality machine translation is available for the given language pair. Our goal was to develop an algorithm working effectively between any European language pairs, and especially between Hungarian and English documents. The Hungarian language has three main obstacles when comparing to other (European) languages: a) loose word order, b) conjugation, c) having a significantly different grammar. These are the reasons – along with small available parallel corpora – that machine translation to and from Hungarian is rather useless for serious applications, often not even understandable by humans, so we decided to go a different path and not to use machine translation in our algorithm.

The new algorithm is based on a distance function between sentences which are evaluated in multiple steps to enable a fast candidate search and a precise comparison between possible translations. It searches for all possible translations, instead of going with one given by an automatic translator. This approach has proved to be effective and eliminated the necessity of using word-sense disambiguation first (at the machine translation stage) and then synonyms in the next step of the system. To show that the algorithm is language independent we included a German-English test corpus as well in addition to the Hungarian-English one.

2. Methodology

The method is based on the understanding that translation is done in most cases on a sentence level: thus sentence chunking is used. Smaller chunks of text (like word n-grams or limbs) often do not correspond between two languages. Larger chunks of text have no distinct borders (if one does not use sentence n-grams); paragraphs can be merged and split easily without hurting the meaning. The plagiarism search is done in three steps: (i) search space reduction, (ii) text similarity evaluation and (iii) post-processing. In the following sections those three steps are described in detail.

2.1. Search Space Reduction

The first step is a standard similarity search, multilingual information retrieval to reduce the search space for the second step. The search space consists – if we consider the English Wikipedia our target – of 200 million chunks, and the best 10-50 candidates will be returned for the next processing step.

The input document is chunked (for details see next chapter) and for each chunk a bag of words is created, filled with all the translations of all the possible lemmas (stems) of all the words of the chunk. Stopwords (including most prepositions) are removed both for reasons of speed and index size reduction, and because most prepositions in English are inflections in some languages like Hungarian, therefore they would have no counterpart in the other language after lemmatization.

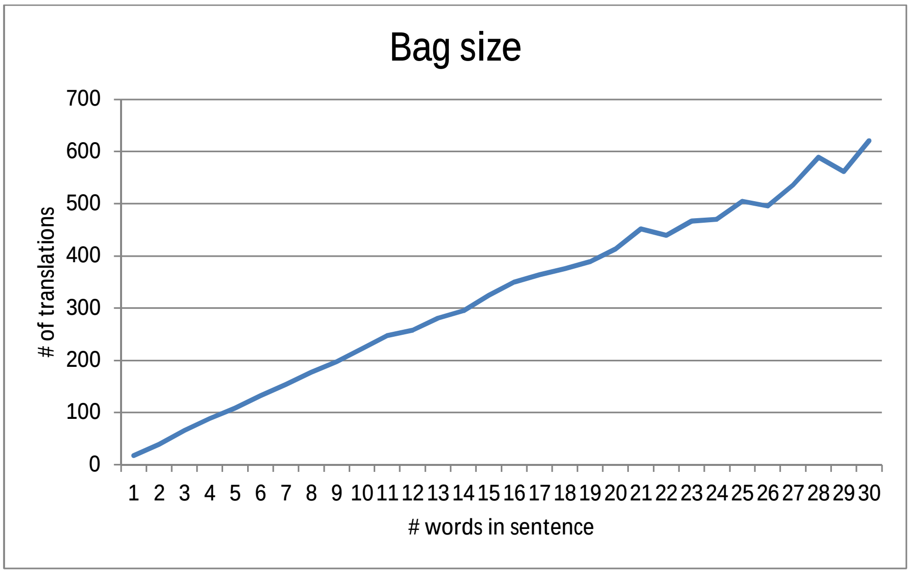


Fig. 1 *Number of translations with respect to sentence length*

By translating an English sentence with the bag of words algorithm into Hungarian, using a dictionary of almost 700 thousand word pairs, the resulting bag is on average 20 times larger than the size of the sentence (see Fig. 1).

2.2. Similarity Evaluation

In this step we calculate a similarity metric (a number) for any two chunks, written in different languages, which is proportional to the similarity between them. The higher the number is, the more similar the chunks are, while negative numbers mean they are not particularly similar.

There are three possible ways to compare texts written in different languages. One is to use non-language-dependent features, like sentence length, which can be useful for text alignment, but is not suitable for comparing texts with unknown content (Gale and Church, 1993). The second is the translation of the texts with an automatic translator to a common language (typically translate one of the texts into the language of the other), which method is used by most current plagiarism detection approaches. (Potthast et al., 2010) The results of machine translation depend largely on the language pair used. Translations between some language pairs and within some domains can be so good that they can be mistaken for a human translation, but some – and sadly Hungarian-English pair is one of them – are of really poor quality (Callison-Burch et al., 2009). The third approach uses language tools to process and compare the two texts. One way of doing this is represented by Ceska et al. where the EuroWordNet thesaurus is used to transform the two texts into a language independent form, which can be then compared. The problem with using a parallel thesaurus is twofold, firstly the number of words are limited compared to a dictionary, e.g. 42-58% of Czech words could be transformed into EWN indexes (Ceska, 2008); secondly and most importantly the necessity of a parallel thesaurus limits the usage to a handful of languages; for Hungarian no such resource is available. In this paper a method is described which uses the third approach, but rather than using parallel thesauri, this one uses dictionaries. Below a similarity metric is defined which is able to calculate the similarity between two texts written in different languages.

Let S denote a sentence of length n, the words of the sentence are represented by w. S_x and S_y are two sentences in different languages.

$$S_x = \{w_{x1}, w_{x2}, w_{x3}, \ldots w_{xn}\}$$

The *trans* function is defined which returns all the translations of a word w

$$\text{if } w_a \in \text{trans}(w_b) \text{ then } w_b \in \text{trans}(w_a)$$

and with the use of that, word identity (≡) is defined for words written in different languages as

$$\text{if } w_a \in \text{trans}(w_b) \text{ then } w_a \equiv w_b$$

To calculate identity, lemmatization is used beforehand on all the words. $| S_x \cap S_y |$ is the number of common words in S_x and S_y where common means $w_{xa} \equiv w_{yb}$.

The similarity metric between two texts (S_x and S_y) is calculated as follows

$$\text{Sim}(S_x, S_y) = \min (| S_x \cap S_y | - | S_x \setminus S_y | , \ | S_y \cap S_x | - | S_y \setminus S_x |$$

The minimum function is needed in order to ensure symmetry and counteract the effects of the two sentences having different number of words (e.g. one containing the other). With that

$$\text{Sim}(S_x, S_y) = \text{Sim}(S_y, S_x)$$

which is expected for translations: if S_x is a possible translation of S_y then S_y has to be also a possible translation of S_x . Another solution would have been to use the $\text{Sim}(S_x, S_y) = | S_x \cap S_y | - | S_x \setminus S_y | - | S_y \setminus S_x |$ but with that, information is lost by smoothing the length difference between the two sentences. Two sentences with 10 and 16 words respectively where all 10 words are common would get the same score as two sentences of 10-10 words each having 8 words in common. The second sentence-pair is a much better translation where the two differing words could result from a missing translation from the dictionary. Our empirical research showed that matching words are more important than missing ones so two constants α and β are introduced to weight the two parts:

$$\text{Sim}(S_x, S_y) = \min (\alpha \cdot | S_x \cap S_y | - \beta \cdot | S_x \setminus S_y | , \ \alpha \cdot | S_y \cap S_x | - \beta \cdot | S_y \setminus S_x |$$

A typical value for α and β are 2 and 1 respectively, meaning matching words are rewarded with 2 points while missing ones are penalized with -1.

This metric can be used to compare two texts to each other. However, as the search space increases linearly with the size of the database, a search space reduction is needed beforehand to select only those candidates from the database which are possible translations of the suspected one.

2.3. Post Processing

The current beta version of the system uses a very quick post processing step. Two thresholds of similarity are defined: Sim_1 and Sim_2 where $Sim_1 < Sim_2$; the current system uses the values 0 and 8 respectively. Two other constants are defined: d_{max} is the maximum distance between chunks and l_{min} is the minimum length of a chunk in words. The following algorithm is used.

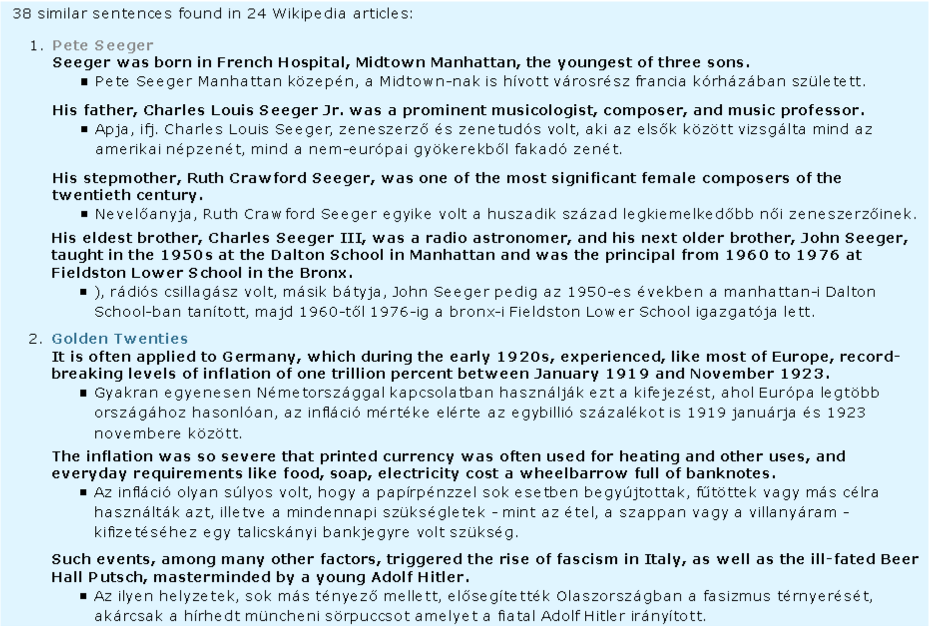
For each chunk all the candidates are kept which have a similarity value above Sim_1 even if there are multiple candidates for a suspected chunk. Hits are sorted by candidate documents. A candidate document is kept and displayed to the user as a hit if

(i) there is at least one chunk for which $Sim > Sim_2$ and its length $l > l_{min}$ (if there is only one short sentence it is disregarded) or if

(ii) there are two similar chunks with chunk positions p_1 and p_2 for which $abs(p_1\text{-}p_2) < l_{min}$ holds true (they are near) or if

(iii) there are more than three chunks.

For each candidate document kept the overall similarity metric SIM is calculated as the sum of all Sim values. Documents are displayed to the user by decreasing SIM values, see Fig. 2 as an example.



38 similar sentences found in 24 Wikipedia articles:

1. Pete Seeger
   Seeger was born in French Hospital, Midtown Manhattan, the youngest of three sons.
   - Pete Seeger Manhattan közepén, a Midtown-nak is hívott városrész francia kórházában született.

   His father, Charles Louis Seeger Jr. was a prominent musicologist, composer, and music professor.
   - Apja, ifj. Charles Louis Seeger, zeneszerző és zenetudós volt, aki az elsők között vizsgálta mind az amerikai népzenét, mind a nem-európai gyökerekből fakadó zenét.

   His stepmother, Ruth Crawford Seeger, was one of the most significant female composers of the twentieth century.
   - Nevelőanyja, Ruth Crawford Seeger egyike volt a huszadik század legkiemelkedőbb női zeneszerzőinek.

   His eldest brother, Charles Seeger III, was a radio astronomer, and his next older brother, John Seeger, taught in the 1950s at the Dalton School in Manhattan and was the principal from 1960 to 1976 at Fieldston Lower School in the Bronx.
   -), rádiós csillagász volt, másik bátyja, John Seeger pedig az 1950-es években a manhattan-i Dalton School-ban tanított, majd 1960-től 1976-ig a bronx-i Fieldston Lower School igazgatója lett.
2. Golden Twenties
   It is often applied to Germany, which during the early 1920s, experienced, like most of Europe, record-breaking levels of inflation of one trillion percent between January 1919 and November 1923.
   - Gyakran egyenesen Németországgal kapcsolatban használják ezt a kifejezést, ahol Európa legtöbb országához hasonlóan, az infláció mértéke elérte az egybillió százalékot is 1919 januárja és 1923 novembere között.

   The inflation was so severe that printed currency was often used for heating and other uses, and everyday requirements like food, soap, electricity cost a wheelbarrow full of banknotes.
   - Az infláció olyan súlyos volt, hogy a papírpénzzel sok esetben begyújtottak, fűtöttek vagy más célra használták azt, illetve a mindennapi szükségletek - mint az étel, a szappan vagy a villanyáram - kifizetéséhez egy talicskányi bankjegyre volt szükség.

   Such events, among many other factors, triggered the rise of fascism in Italy, as well as the ill-fated Beer Hall Putsch, masterminded by a young Adolf Hitler.
   - Az ilyen helyzetek, sok más tényező mellett, elősegítették Olaszországban a fasizmus térnyerését, akárcsak a hírhedt müncheni sörpuccsot amelyet a fiatal Adolf Hitler irányított.

Fig. 2 *Example extract from the result displayed to the user*

3. Findings

Three main evaluations were done: the first about the ideal dictionary size, as this influences search time; the second is the evaluation of the new similarity metric; the third and final is the evaluation of the system as a whole.

3.1. Optimal Dictionary Size

As discussed above, the size of the dictionary increases not only the probability that two translations will be found similar, but also increases the bag size and the runtime accordingly. The dictionary used by the system had no information about which translations are the most common for a given word, so a separate, simple list of words (lemmas) by frequency was used to limit the number of translations of a word to the tr_{max} most common ones. Please note that this is not the ideal solution as word frequency in a language does not indicate the right order of translation for all words; nonetheless, it was the best possible approximation.

With that combined dictionary a series of tests was executed, where the maximum number of translations per word was limited to between 1 to 100 words. The translations were ordered by the occurrence frequency in their respective language, from the most frequent to the least frequent one.

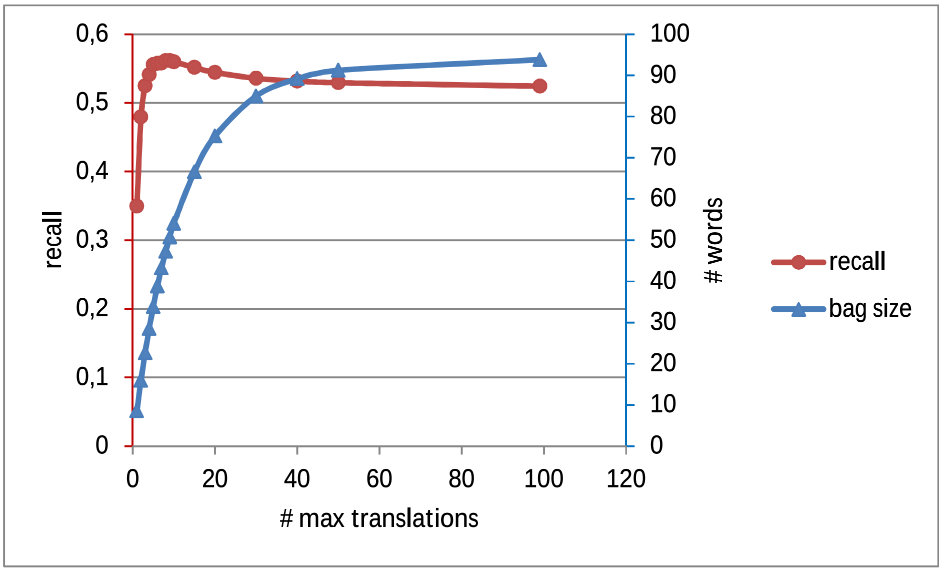


Fig. 3 *Number of words and recall as a function of the number of translations per word*

Fig. 3 shows the correlation of the recall and the number of words in the bag (which is linear to the query speed) as a function of the maximum number of translations used. It is clearly visible that the recall has a maximum; there is an optimal number of translations that can be used, and after which the curve declines. By enlarging the corresponding part (see Fig. 4) it is visible that the maximum is at 8 translations but from 5 translations on the difference is negligible. Based on this finding, and because the bag size grows considerably between 5 and 8 (from 33.9 to 47.3), the parameter $tr_{max} = 5$ is used for the final system and for all further tests, if not stated otherwise.

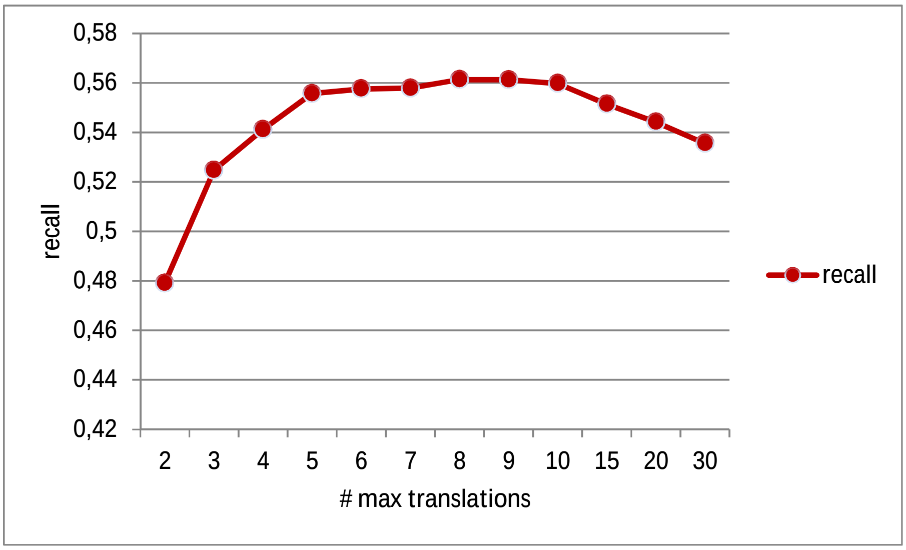


Fig. 4 *Recall as a function of the number of translations per word*

3.2. Similarity Metric Evaluation

The similarity metric was tested separately on two human translated parallel corpora, the Hunglish (Tóth et al., 2008) and SzegedParalell (Varga et al., 2005), incorporating 1.3 million and 100 thousand sentences respectively. Hunglish is a large collection created from document pairs by automatic methods. “Sometimes parts of the documents are not in perfect correspondence, due to liberal translation, or even skipping of some segments by the translator. These may lead to erroneous sentence pairs.” (Vargaet al., 2005) SzegedParalell is much smaller in size but was checked and corrected manually by its creators.

For our tests recall is calculated as the probability of two chunks which are translations of each other to get a Sim score higher than Sim_1 (as those are kept in the post processing). The same value for Sim_2 is also calculated for demonstration purposes.

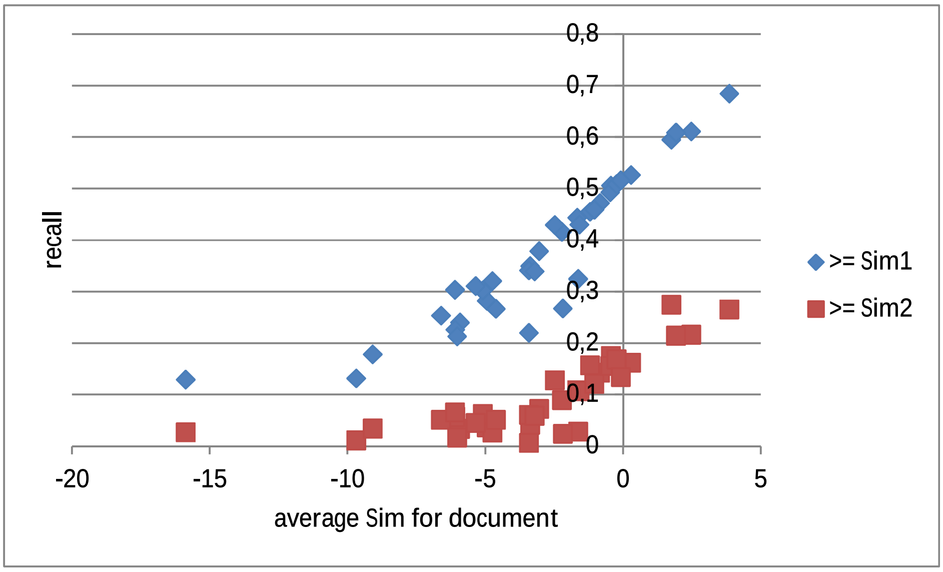


Fig. 5 *Recall values for Sim1 and for Sim2 as a function of average Sim*

Fig. 5 shows the recall values for Sim1 and for Sim2 thresholds in the function of the average Sim value for 37 documents from Hunglish. This was an expectable result, a higher average Sim value results in a higher recall.

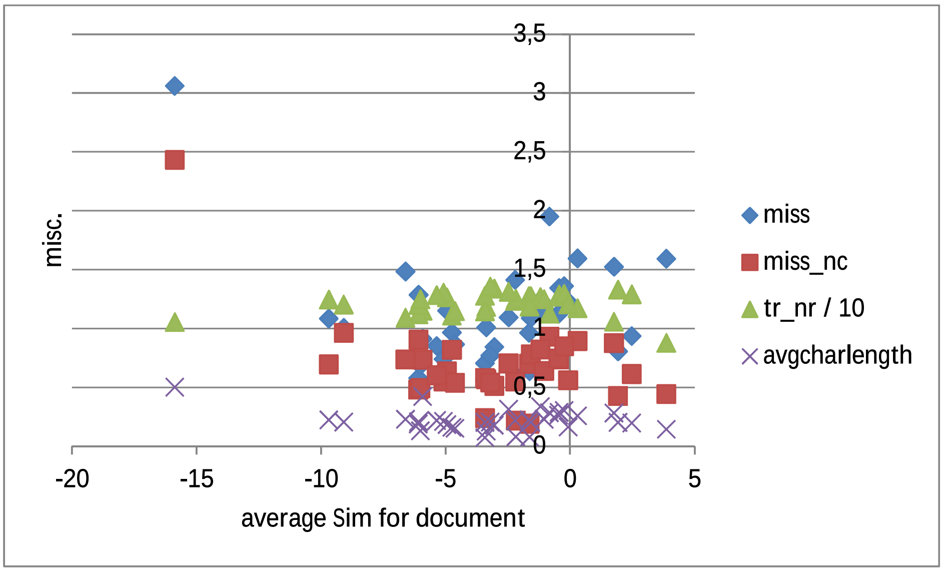


***Fig. 6** Four other document indicators as a function of average Sim*

Four other document indicators were tested whether they correlate to the average Sim score (or to the recall): average missing words from dictionary per sentence (mis), average missing words from dictionary per sentence without the capital words (mis_nc), average number of translations per word (tr_nr), and average length of a sentence in characters (see Fig. 6). They do not correlate, if they had correlated they could have been used to indicate in advance the quality of the search. It is obvious that if a lot of words to be translated are missing from the dictionary the search will be of poor quality; but with a large dictionary no difference could be found even between a text from Hans Christian Andersen and the Bible.

	# pairs	>= Sim$_1$ (recall)	>= Sim$_2$
Hunglish	1 297 696	0,51	0,10
SzegedP.	99 345	0,52	0,17
WP ger-eng	66 645	0,65	0,29

***Table 1** Recall values for different corpora*

Table 1 summarizes the results for the different parallel corpora, recall here means the average recall for one single chunk, not a larger overlap. As it can be seen, the machine translated corpus has a better recall value, which could result either from the fact that the two other corpora include a lot of old texts and literary translations, or that the machine translation produces a poorer quality text.

3.3. System Evaluation

For the evaluation of the whole system the English Wikipedia and a randomly selected smaller corpus has been used, containing 65 000 parallel sentences from Wikipedia in the English original, translated Hungarian (WP hun-eng) and translated German (WP ger-eng). Machine translation was done with Google Translate API. Two dictionaries, the English-Hungarian with 700 thousand word-pairs, and the English-German with 150 thousand word pairs were used. The system was also tested on a very small – 100 sentences long – hand translated Hungarian corpus from the Wikipedia. Its result was similar to that of the machine translated, but as the corpus was so small there is a great uncertainty, therefore the results are not presented here. It is important to note that to use machine translation for the tests is only possible because the algorithm does not depend on any automatic translation. To use a large hand translated corpus would have been desirable, but such a corpus was not available.

The English Wikipedia was uploaded into the database, which as of this writing consists of 3.8 million articles, about 200 million chunks. At the search space reduction level this database was queried with the translated bag of words and the first 50 results were evaluated. The research showed (see Fig 7.) that even the first 10 would be enough as there is no significant difference, the right sentence is with 0.44 probability the first one, 0.13 that it is between 2nd and 10th, there is only 0.06 probability that it is between the 10th and the 50th (and 0.37 that it is not within the first 50 at all). For the German corpus (because of the smaller dictionary) these numbers were a little bit different: 0.34, 0.11, 0.06 and 0.49, respectively.

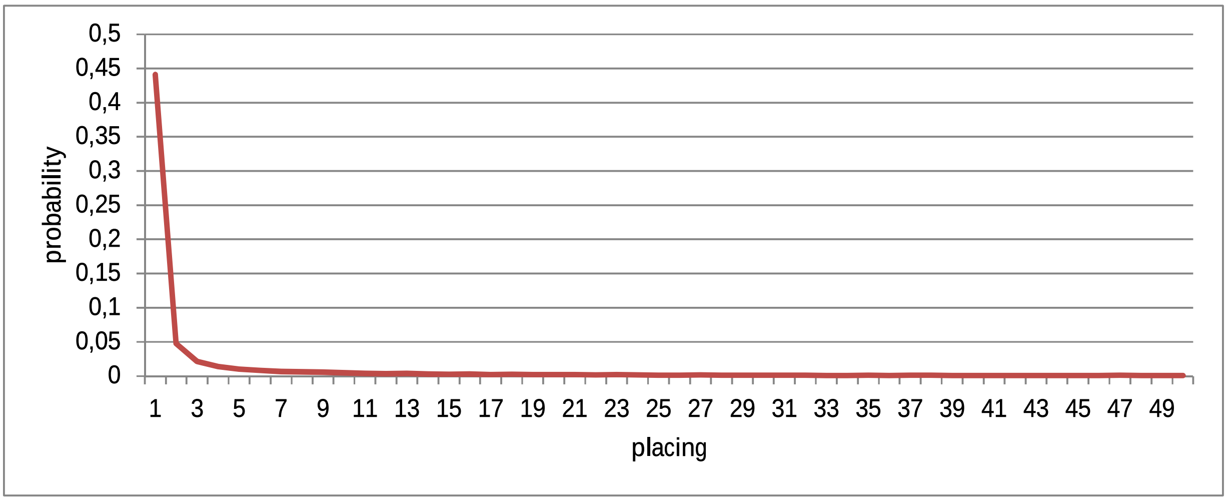


Fig. *7 Placing of the right chunk at the query level (hun-eng)*

This missing 0.49 can be seen on the end results, and is mostly caused by the short sentences as shown in Fig. 8. The recall increases with the length of the suspected sentences (counted in number of words): the dotted curve represents the similarity metric run alone on the German-English parallel corpus (simulating the effects of a perfect search space reduction), the continuous curve is achieved by querying the same sentences from the system which was uploaded with the full English Wikipedia; hits are clearly lost here due to the information retrieval method used. The Hungarian-English query shows exactly the same result.

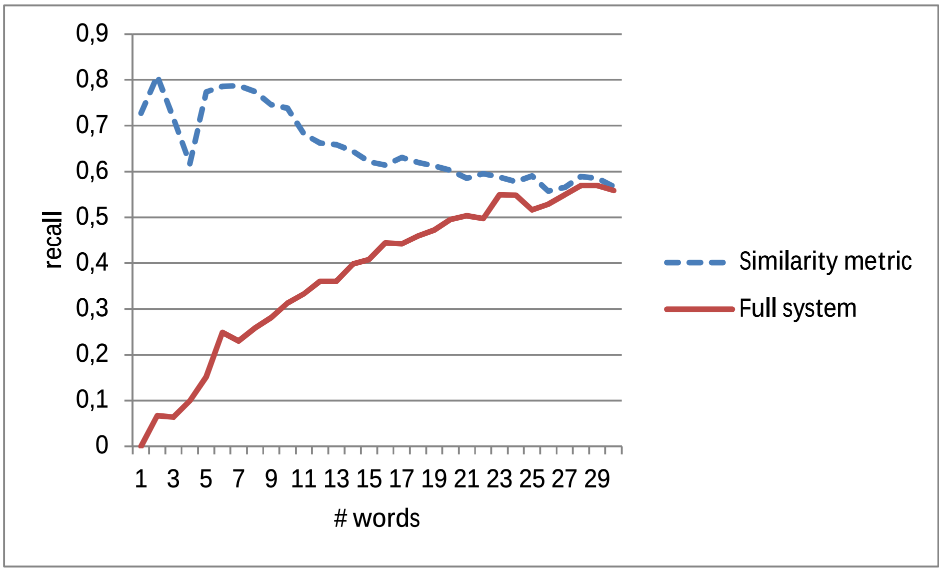


Fig. 8 The recall as a function of sentence length

4. Discussion

With the use of the new Similarity Metric, the results returned by the query could be filtered out effectively, see Fig. 9 (the origin is not at 0), where probability is drawn that a translation is correctly recognized and placed among the first *n* places. There is exactly 10-percent-point increase in the first place in favour of the Similarity Metric, and the steep curve shows that translations can be distinguished from other sentences, but it is also obvious that 3 per cent of the results are missed.

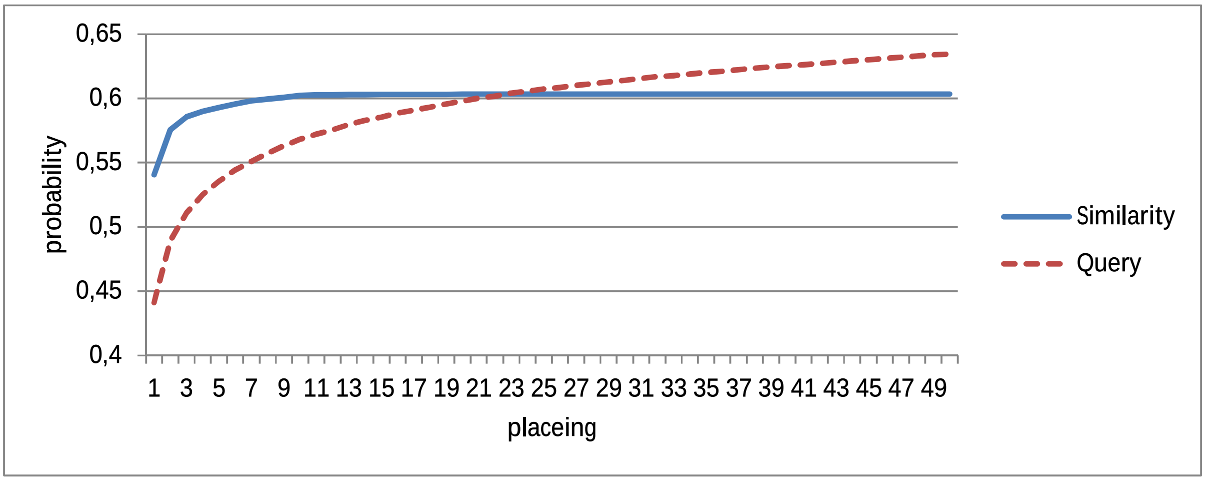


Fig. 9 *Probability of being ranked above a certain value*

The curve above shows the probabilities that a translated sentence is correctly found and ranked first, but the Post Processing step does not use the ranks, but the scores, which shows a similar result (see Table 2).

	1. rank	1-10. ranked	>= Sim_1	>= Sim_2
WP hun-eng	0,54	0,60	0,62	0,52
WP ger-eng	0,31	0,34	0,40	0,23

Table 2 *Ranks and recall of the two Wikipedia corpora*

Until now the calculations were done using single chunks of a document. As the goal of most plagiarism checkers is not to find an overlap of one or two sentences, but to effectively protect against large-extent copies or translations we have to scale up our results. Assumed that the probability to find a sentence is independent from the previous one, Table 3 summarizes the probabilities to find at least *x* out of *y* sentences, calculating with the Sim_1 value for the German-English corpus.

y	***x=1***	***x=2***	***x=3***	***x=4***	***x=5***
1	0,4				
2	0,64	0,16			
3	0,784	0,352	0,064		
4	0,8704	0,5248	0,1792	0,0256	
5	0,92224	0,66304	0,31744	0,08704	0,01024
6	0,953344	0,76672	0,45568	0,1792	0,04096
7	0,972006	0,84137	0,580096	0,289792	0,096256
8	0,983204	0,893624	0,684605	0,405914	0,17367
9	0,989922	0,929456	0,768213	0,51739	0,266568

10	0,993953	0,953643	0,83271	0,617719	0,366897
11	0,996372	0,969767	0,881083	0,703716	0,467226
12	0,997823	0,980409	0,916557	0,774663	0,561822
13	0,998694	0,987375	0,942098	0,83142	0,646958
14	0,999216	0,991902	0,960208	0,875691	0,720743

Table 3 *Probabilities to find at least x number of y number of sentences*

5. Conclusion and Future Plans

We showed a possible alternative method to using machine translation for cross-language plagiarism detection: the use of information retrieval method and a new cross-language similarity metric. The algorithm is capable of detecting a 10-sentence long translation with over 95% probability for German-English and 99% for Hungarian-English language pair, noted that this was tested on a machine translated corpus.

The evaluation of the new algorithm has not yet been finished, however, at the conference we would like to be able to present our findings with obfuscated translations. The precision measures did not produce relevant output, as there are too many duplicate contents in the database which were detected as false positives. This has to be tested with an artificial test corpus.

The bottleneck of the system is the search space reduction, at a later stage that algorithm could also be revised, but this was not within the scope of this research. The similarity metric can also be extended by recognizing phrases and expressions. By using a professional dictionary with the actual translation frequency of a given word the speed or the quality of the system could be boosted. Using a POS tagger could enable the system to weight the words according to their importance, e.g. a noun is most probably more important than a preposition, and numbers could also be weighted differently.

The human tests showed that the quality of the returned result is more than adequate to be used in a production environment, therefore this algorithm was integrated into our online plagiarism search tool at the end of 2011 and other languages were added to it as well. The aim was to present that the algorithm itself is language independent; only the pre- and post-processing steps require some knowledge about the language and grammar of the text. Consequently, the same algorithm can be used to detect plagiarism within one language and – being integrated into the system – is currently able to compare English, German and Hungarian texts to the English and Hungarian Wikipedia. However, this spectrum of languages will soon widen, as new languages and databases are added to the system.

6. Acknowledgements

I would like to thank Miklós Vajna for his work in the Wikipedia converter, András Kornai and Gábor Recski for their help in NLP related questions, András Micsik for the pre-review and the many valuable suggestions. Special thanks to Éva Virág for the proofreading of the text.

7. References

Bailey, J., 2011. *PlagAware Takes Top Honors in Plagiarism Checker Showdown.* [online] Available at: http://www.plagiarismtoday.com/2011/01/13/plagaware-takes-top-honors-in-plagiarism-checker-showdown/ [Accessed 30 March 2012].

Callison-Burch, C., Koehn, P., Monz, C. and Schroeder, J., 2009. *Findings of the 2009 Workshop on Statistical Machine Translation*. EACL Workshop on Statistical Machine Translation. pp. 1-28. Association for Computational Linguistics, Athens, Greece.

Ceska, Z., Toman, M. and Jezek, K., 2008. *Multilingual Plagiarism Detection*. Proceedings of the 13th International Conference on Artificial Intelligence. pp. 83–92. Springer Verlag, Berlin Heidelberg

Copy, Shake, and Paste blog, 2012. *A blog about plagiarism from a German professor, written in English.* [online] Available at: http://copy-shake-paste.blogspot.com/ [Accessed 30 March 2012].

Copyscape, 2012. *Searches for copies of your webpage on the web.* [online] Available at: http://www.copyscape.com/

Gale, W. A. and Church, K. W., 1993. *A Program for Aligning Sentences in Bilingual Corpora.* Computational Linguistics 19 (1), pp. 75–102

PlagAware, 2012. *Protecting a website's content against content theft and for performing plagiarism assessments on texts.* [online] Available at: http://www.plagaware.com/

Plagiarisma, 2012. *Free online plagiarism checker.* [online] Available at: http://plagiarisma.net/

Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein B. and Rosso, P., *Overview of the 2nd International Competition on Plagiarism Detection* [online] Available at: http://www.clef2010.org/resources/proceedings/clef2010labs_submission_125.pdf [Accessed 30 March 2012].

Storm, C., 2010. *Translated and Paraphrased Plagiarism*. Fourth International Plagiarism Conference. Newcastle upon Tyne, UK, 2010

Tóth, K., Farkas, R. and Kocsor, A., 2008. *Hybrid algorithm for sentence alignment of Hungarian-English parallel corpora.* Acta Cybernetica 18(3), pp. 463-478. Available at: http://www.inf.u-szeged.hu/rgai/corpus_paralell [Accessed 22 Feb 2012].

Turnitin, 2012. *Academic plagiarism detector*. [online] Available at: https://turnitin.com/

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V. and Nagy V., 2005. *Parallel corpora for medium density languages*, In Proceedings of the RANLP 2005, pages 590-596. Available at: http://szotar.mokk.bme.hu/hunglish/search/corpus [Accessed 22 Feb 2012].

Weber-Wulff, D., 2010. *Results of the Plagiarism Detection System Test 2010*. [online] Available at: http://plagiat.htw-berlin.de/software-en/2010-2/ [Accessed 30 March 2012].