# A light way to collect comparable corpora from the Web

**Ahmet Aker**$^{\star}$**, Evangelos Kanoulas**$^{\dagger}$**, Robert Gaizauskas**$^{\star}$

$^{\star}$Department of Computer Science, University of Sheffield, Sheffield, UK
$^{\dagger}$Google Research, Zurich, Switzerland
Email: a.aker@dcs.shef.ac.uk, ekanou@google.com, r.gaizauskas@dcs.shef.ac.uk

### Abstract

Statistical Machine Translation (SMT) relies on the availability of rich parallel corpora. However, in the case of under-resourced languages, parallel corpora are not readily available. To overcome this problem previous work has recognized the potential of using comparable corpora as training data. The process of obtaining such data usually involves (1) downloading a separate list of documents for each language, (2) matching the documents between two languages usually by comparing the document contents, and finally (3) extracting useful data for SMT from the matched document pairs. This process requires a large amount of time and resources since a huge volume of documents needs to be downloaded to increase the chances of finding good document pairs. In this work we aim to reduce the amount of time and resources spent for tasks 1 and 2. Instead of obtaining full documents we first obtain just titles along with some meta-data such as time and date of publication. Titles can be obtained through Web Search and RSS News feed collections so that download of the full documents is not needed. We show experimentally that titles can be used to approximate the comparison between documents using full document contents.

**Keywords:** Comparable Corpora, News Search, Statistical Machine Translation

## 1. Introduction

Statistical Machine Translation (SMT) relies on the availability of rich parallel resources. However, often parallel resources are not readily available for under-resourced languages or specific narrow domains. This leads to underperforming machine translation systems. To overcome the low availability of parallel resources the machine translation community has recognized the potential of using comparable resources as training data (Rapp, 1999; Munteanu and Marcu, 2002; Sharoff et al., 2006; Munteanu and Marcu, 2006; Kumano et al., 2007; Barzilay and McKeown, 2001; Kauchak and Barzilay, 2006; Callison-Burch et al., 2006; Nakov, 2008; Zhao et al., 2008; Marton et al., 2009).

A critical first problem with such an approach is actually identifying and gathering corpora with the potential for improving SMT systems. Attempts at gathering comparable corpora from the Web have been made (Braschler, 1998; Resnik, 1999; Huang et al., 2010; Talvensaari et al., 2008). The process of obtaining such corpora involves (1) downloading for each language a separate set of documents, (2) matching documents between the two languages by comparing document contents, and finally (3) extracting useful units for SMT from the matched document pairs by applying approaches such as that described in Munteanu and Marcu (2006). In this work we focus on steps 1 and 2. For step 1 past studies (see Section 2.) commonly use a cross-language information retrieval (CLIR) approach or some rules of thumb, where source and target language document collections are gathered by monolingual crawling. Step 2 is performed using weak translation methods and search to pair target language documents with those in the source collection. These steps are resource intensive and time consuming requiring huge amounts of disk space and fast computing machines.

In this work we aim to reduce the amount of time and resources required by steps 1 and 2. We collect comparable corpora from the Web by focusing only on news articles. When collecting comparable corpora we only rely on the title of the documents and do not use the article contents. Using only the title for comparison clearly saves time and computation, since we obtain the titles through Google News Search and RSS News feed collections, so that downloading of the full document contents prior to determining comparability is not needed. In this paper we show experimentally that the title can be used to successfully match the documents instead of using the full document contents. The full content of the matched pairs can then be downloaded.

Some news stories run for some time and the initial report gets developed in follow up articles. In many of the follow up news articles a similar title to the earlier one(s) is used, although the content of the follow up news starts diverging from the original report. For this reason we also combine the title with publishing date and time which are also available through the Google News Search and RSS News feeds to investigate their contribution to the quality of the document pairs.

Through our study we build a framework for obtaining comparable corpora for various language pairs, such as *English-German, -Greek, -Croatian, -Estonian, -Latvian, -Romanian, -Lithuanian* and *-Slovenian*. In this paper, we focus on *English-German* and *English-Greek* with the latter being an under-resourced language. We evaluate the collected corpora by asking human assessors to assess the level of their comparability.

## 2. Related Work

Constructing comparable corpora has been investigated in earlier studies. Braschler (1998) uses existing news document collections in English-German and English-French and investigates different ways to align the most potentially useful pairs across monolingual collections by using proper nouns, numbers, date similarity and content bearing words. Munteanu et al. (2004) and Munteanu and Marcu (2005) use dictionaries trained using initial parallel data to create comparable corpora. For every document in the source language in the comparable corpora, the top five dictionary based translations of every word are used to create a query to search all documents in the target language. This search is limited to articles published within 5 days of the source text and only top 20 ranked articles are returned and paired with the source text. More recently, Huang et al. (2010) describe methods for obtaining comparable corpora for English-Chinese documents using Cross Lingual Information Retrieval (CLIR) techniques. Monolingual documents are crawled from manually selected websites and later paired using CLIR. Another CLIR approach is described by Talvensaari et al. (2008). The authors obtain comparable corpora for English, Spanish and German. They manually select a set of topic words and then use them in a monolingual crawler. After the monolingual collections are downloaded, CLIR is used to pair the documents across languages based on word co-occurrences in the source and target documents. Pouliquen et al. (2004) focus on clustering news articles. The authors first cluster news articles published on the same day monolingually using content bearing words and named entities. Then they map the news articles within the same cluster to a multilingual thesaurus from which a list of concept terms is extracted (each concept term has unique translations in 22 different languages). The thesaurus concept terms are then used to link news cluster written in different languages. Similar to Pouliquen et al., the work by Montalvo et al. (2006) also makes use of named entities to pair documents.

In summary, in all cases the full document content is required to make judgements about the comparability level of two documents. However, to do this a huge number of documents in different languages must first be downloaded from the web and preprocessed. Downloading documents and preprocessing them is time and resource intensive. In addition, only a small portion of these downloaded documents will be included in the comparable corpora. The vast majority of the documents will be discarded.

We aim to overcome this problem and create comparable corpora by just using document or article titles. Obtaining such titles is less resource and time consuming. We focus on news articles and thus make use of the fact that titles in news articles are a good indicator for the content of the document. In a text summarization framework Edmundson (1969) scored sentences which contain words from the title higher than sentences which do not include these terms. The motivation behind this was that writers re-use the words from the titles in the subsequent sentences when they write their articles. Lopez et al. (2011) have analyzed 300 titles of news articles and showed that 66% of the title words occur in the articles. Therefore we use titles as representatives of document contents and use them to judge documents comparability level.

## 3. Collecting News Titles

In order to collect comparable corpora we first collect news article titles through Google News Search and RSS News feeds. We only download current news articles and do not search for articles in news archives or on the entire Web. Searching in a bigger space causes more noise in the pairing process (Section 3.1.) than when the focus is only on current news.

This title collection should be of high recall, i.e. contain as many potentially useful titles as possible. Thus, during this initial process of collecting our "working material" we only care about recall and ignore precision (the proportion of the collected titles that are actually comparable with each other). After collecting these general titles, we apply different heuristics to pair them. To collect the title corpora we adopt the following process:

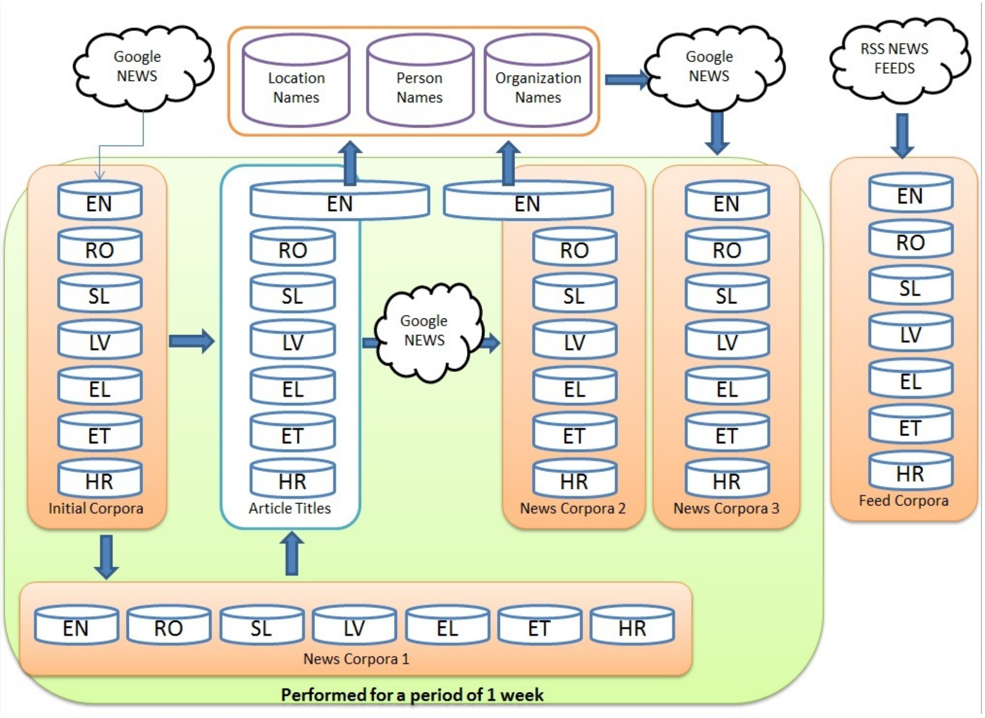
1. We first collect *initial corpora* of titles from news article monolingually using Google News. For each language we iteratively download titles from news in different topic categories, such as *economics, world, politics*, etc. We set the iteration time to 15 minutes. Apart from the title for each search result we also have information about the *date and time of publication*, the *url* to the actual article and another url that it is used by Google News to show all related articles about the same topic in a cluster – we refer to this url as *cluster url*. We refer to the titles obtained by this first step as the *initial corpora* (Figure 1).

2. We make use of the Google News clustering of News articles that are found to be similar to each other and for each title in our *initial corpora* we collect titles of articles that are clustered with it. More precisely we follow the *cluster url* and download the first 30 articles from the cluster. We refer to these corpora as *news corpora 1*. Clearly, following these two steps one can collect as many titles as one wants spanning a period of time. In our case this period was a week. In this way we always download the current news. This means if our method runs, e.g. for one week, the first downloaded news article will be one week old.

3. We then use the titles from the *initial corpora* and *news corpora 1* as queries and perform a monolingual Google News search. We extract the titles from the search results and these constitute *news corpora 2*. When performing this search we restrict the date of the search to a maximum of one week from the moment the search is performed. Furthermore, we collect *news corpora 2* in parallel with the *initial* and *news 1* corpora. As shown in Figure 1 we run these processes for a week.

4. Next, we further expand the collection of article titles to include *news corpora 3*. For this, we take the article titles from the *initial corpora, news corpora 1* and *news corpora 2* for the English collection only. We parse them for named entities such as *person, location*

Figure 1: Crawl Steps.



and *organization* names.[1] For each named entity type we do the following: we translate the entities into the language in which the search will be performed (using Google Translate) and perform a Google News Search using the translated entity as a query. The search is restricted to a maximum of one week prior to the publication date of the article.

5. Finally, Google News does not support all languages equally. Languages such as German or Greek are well supported by Google News, i.e. articles of different news agencies are preprocessed and listed by Google News. However, this is not the case for languages such as Latvian, Lithuanian, Estonian, etc. Due to this fact there is a data scarcity problem in those languages. To overcome this problem we also manually identify a good number of RSS News feeds for each language from which we extract similar information as in the Google News Search.

### 3.1. Document Alignment

In the alignment phase the goal is to match the article titles from the collected corpora and only download the actual article contents for those matching pairs to obtain a comparable corpus.

Matching news by title similarity (*TS*) is performed by computing the cosine similarity across the titles' term frequency vectors. Thus, each title pair is scored between 0 and 1. Before computing the cosine measure we also ensure that both titles (after removing the stop words) have at least 5 content words on both sides. We have experimentally observed that a news title with at least 5 content words is best to represent the actual document content. We translate the foreign title into English using Google Translate. We also combine *TS* with the following heuristics to investigate their impact on the quality of the produced pairs:

- *HS*: Each article title pair is scored by $\frac{1}{h+1}$, where $h$ is the time difference in hours, with $h \in [0, ..., 23]$. Articles published within the same hour get a score of 1. If the time difference is greater than 23h then HS is set to 0.
- *DS*: We score each article title pair by the publishing date difference between the two articles ($\frac{1}{d+1}$, where $d$ is the date difference, with $d \in [0, ..., 7]$). Articles published on the same date get a score of 1. We set *DS* to 0 when the publishing date is greater than 7 days.
- *TLD*: We score each article title pair by $\frac{1}{w+1}$, where $w$ is the difference in content word count (starting from 0). Article titles with the same length get a score of 1.

## 4. Evaluation

We create different combinations of the heuristics and evaluate the quality of the results. We use a linear combination of each heuristic with equal weight. Each combination produces a ranked list of article title pairs. The following list summarizes the different heuristic combinations:

- *TS*: Title cosine similarity.
- *TS_HS*: Title cosine similarity and time difference.
- *TS_DS*: Title cosine similarity and date difference.
- *TS_TLD*: Title cosine similarity and title length difference.
- *TS_TLD_HS*: Title cosine similarity, title length difference and time difference.

[1] For named entity parsing we use OpenNLP tools: http://incubator.apache.org/opennlp/

Table 1: Ranking correlation between the different heuristic combinations for the English-German pairs.

| | TS | TS_HS | TS_DS | TS_TLD | TS_TLD_HS | TS_TLD_DS |
|---|---|---|---|---|---|---|
| TS | – | 1 | 0.94 | 1 | 0.99 | 0.73 |
| CS | 0.23 | 0.19 | 0.15 | 0.09 | 0.16 | 0.17 |

Table 2: Ranking correlation between the different heuristic combinations for the English-Greek pairs.

| | TS | TS_HS | TS_DS | TS_TLD | TS_TLD_HS | TS_TLD_DS |
|---|---|---|---|---|---|---|
| TS | – | 1 | 0.82 | 1 | 0.95 | 0.78 |
| CS | 0.11 | 0.21 | 0.14 | 0.18 | 0.25 | 0.25 |

- *TS_TLD_DS*: Title cosine similarity, title length difference and date difference.

We perform a ranking comparison between the different ranked lists of title pairs and human assessment on the aligned articles.

### 4.1. Evaluation: Ranking Order

We compare the quality of the pairs produced by the different heuristic combinations with the ones obtained when the article content is used. To compute the content similarity, first we consider the union of the top 1K pairs of titles ranked by each one of six aforementioned methods. The maximum number of pairs in the union is 6K. Following the corresponding URLs we download content (text) of the article pairs and compute the cosine similarity over term frequency vectors of the entire article. We use an HTML parser[2] to extract text from the HTML documents. Before comparing the article contents, each foreign article text is translated into English using Google Translate. The comparison of article texts produces another ranked list of article pairs which we refer as the *CS* list.

We compared the rankings of each similarity heuristic using Kendall's $\tau$. Kendall's $\tau$ values close to 1 reflect rankings very similar to each other, while values very close to 0 reflect independent rankings. The results are shown in Table 1 and 2. As one can observe, in both German and Greek, the results in the first row show that the rankings produced by different heuristic combinations correlate very highly with the original title similarity. Thus, date, time and title length do not dramatically change the matching process. On the other hand the correlation between *CS* and the other heuristic combinations is rather low as shown in the second row of both Tables 1 and 2. Thus, using the title (along with other meta-data) does not produce the same matches as when using the entire article. The next step is to investigate how humans judge the different rankings produced for the two cases (title similarity and meta-data versus content similarity).

### 4.2. Evaluation: Human Judgment

In the human evaluation we asked assessors to judge the comparability of each aligned document pair. We use five comparability classes proposed by Braschler (1998): *same story, related story, shared aspect, common terminology and unrelated* to judge each document pair manually. We hypothesize that if two news articles are about the "same story" then it is more likely that they contain useful fragments for SMT than if they are "unrelated". The document contents were shown to the assessors side-by-side. The design of the assessment implementation is shown in Figure 2.

Table 3: English-German document pair evaluation results. Results of both assessors are taken together. The numbers are percentage values.

| | same story | related story | shared aspect | common terminology | unrelated |
|---|---|---|---|---|---|
| TS | 74 | 24 | 2 | 0 | 0 |
| TS_HS | 88 | 12 | 0 | 0 | 0 |
| TS_DS | 76 | 18 | 6 | 0 | 0 |
| TS_TLD | 74 | 24 | 2 | 0 | 0 |
| TS_TLD_HS | 86 | 12 | 2 | 0 | 0 |
| TS_TLD_DS | 72 | 22 | 6 | 0 | 0 |
| CS | 75 | 21 | 4 | 0 | 0 |

Table 4: English-Greek document pair evaluation results. Results of four assessors are taken together. The numbers are percentage values.

| | same story | related story | shared aspect | common terminology | unrelated |
|---|---|---|---|---|---|
| TS | 50 | 12 | 24 | 7 | 7 |
| TS_HS | 56 | 15 | 20 | 5 | 4 |
| TS_DS | 62 | 8 | 30 | 0 | 0 |
| TS_TLD | 50 | 8 | 25 | 11 | 6 |
| TS_TLD_HS | 70 | 8 | 20 | 2 | 0 |
| TS_TLD_DS | 42 | 18 | 32 | 8 | 0 |
| CS | 29 | 19 | 32 | 6 | 14 |

We employed a "pooling" approach similar to the one used in TREC[3] and ImageCLEF[4], and constructed a depth-30 pool by considering the union of the top 30 document pairs coming from each one of the approaches under consideration: *TS, TS_HS, TS_DS, TS_TLD, TS_TLD_HS, TS_TLD_DS and CS.*

The document pairs in the pools for the two languages were shown to two native German and eight native Greek speakers respectively. All judges were also fluent in English. For German each participant judged all the pairs in the German pool. In case of the Greek experiment each quarter of the pool was shown to two different assessors.

From the results shown in Tables 3 and 4 we can see that the documents aligned with the title and meta-data information are mainly judged as being "same story" and "related story". For English-German the best performance is achieved when the title similarity is combined with the publishing time ($TS\_HS$). In case of the English-Greek pairs we can see that *HS* plays also an important role.

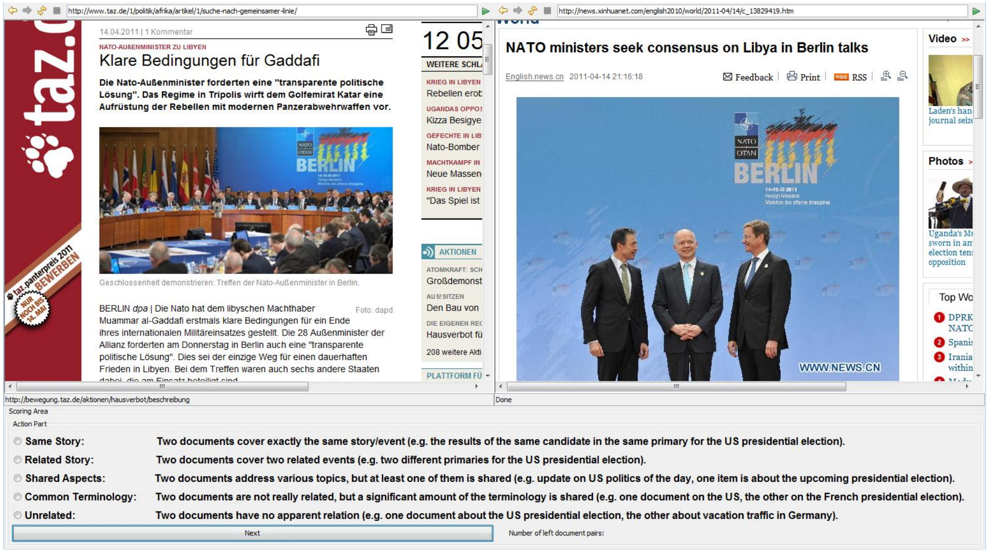
The reason for the positive impact of *HS* may be that it reflects the way news events emerge. Two news articles published very close to each other in time are likely to report the same news event in the same way. However, over time a news event develops and changes so any new report about it will differ from the first reports. Although the new reports are also about the same general event, the contents differ from the first reports and become reports of related stories or reports which share only some aspects with the first ones.

[2] Boilerpipe – http://code.google.com/p/boilerpipe/ – is used to extract the textual content from the URL

[3] http://trec.nist.gov

[4] http://www.imageclef.org

Figure 2: Evaluation Tool.



http://www.taz.de/1/politik/afrika/artikel/1/suche-nach-gemeinsamer-linie/

14.04.2011 | 1 Kommentar

NATO-AUßENMINISTER ZU LIBYEN

Klare Bedingungen für Gaddafi

**Die Nato-Außenminister forderten eine "transparente politische Lösung". Das Regime in Tripolis wirft dem Golfemirat Katar eine Aufrüstung der Rebellen mit modernen Panzerabwehrwaffen vor.**

Geschlossenheit demonstrieren: Treffen der Nato-Außenminister in Berlin.

Foto: dapd

BERLIN *dpa* | Die Nato hat dem libyschen Machthaber Muammar al-Gaddafi erstmals klare Bedingungen für ein Ende ihres internationalen Militäreinsatzes gestellt. Die 28 Außenminister der Allianz forderten am Donnerstag in Berlin auch eine "transparente politische Lösung". Dies sei der einzige Weg für einen dauerhaften Frieden in Libyen. Bei dem Treffen waren auch sechs andere Staaten

http://bewegung.taz.de/aktionen/hausverbot/beschreibung

http://news.xinhuanet.com/english2010/world/2011-04/14/c_13829419.htm

NATO ministers seek consensus on Libya in Berlin talks

English.news.cn 2011-04-14 21:16:18

Done

Scoring Area

Action Part

○ Same Story: Two documents cover exactly the same story/event (e.g. the results of the same candidate in the same primary for the US presidential election).

○ Related Story: Two documents cover two related events (e.g. two different primaries for the US presidential election).

○ Shared Aspects: Two documents address various topics, but at least one of them is shared (e.g. update on US politics of the day, one item is about the upcoming presidential election).

○ Common Terminology: Two documents are not really related, but a significant amount of the terminology is shared (e.g. one document on the US, the other on the French presidential election).

○ Unrelated: Two documents have no apparent relation (e.g. one document about the US presidential election, the other about vacation traffic in Germany).

This fact is supported by the results shown in Tables 3 and 4 where we see that any combination of heuristics without *HS* has higher "shared aspect" than the combinations with *HS*. The heuristic *DS* is also meant to capture news articles about the same story. However, since *DS* uses day level difference in scoring, it can only achieve similar performance to *HS* for stories which do not emerge very quickly.

For English-Greek we get the best results when *TS* and *HS* are combined with *TLD* ($TS\_TLD\_HS$) – note that adding *TLD* to $TS\_HS$ in English-German leads to almost as good results as those obtained with $TS\_HS$ only. In general the heuristic *TLD* plays also a role in the title method. It ensures that titles with no length difference are scored higher than the ones which vary a lot in length. We computed the average title length difference for each language.[5] The English titles contain on average 6.8 content words, the German titles 6.5 and the Greek titles 5.8. These figures show that the English and Greek titles vary from each other more than the English and German ones. We think that this may explain why *TLD* has more impact on the English-Greek results than it has on the English-German ones.

In the ranking results shown in Tables 1 and 2 we see that there is no correlation between the ranked list of article pairs produced by *CS* and the article rankings of the other heuristics. However, from the results shown in Tables 3 and 4 we see a different picture. In the case of the German-English pairs the title similarity heuristics perform as well or better than the *CS* measure, while for the English-Greek pairs title similarity heuristics perform significantly better than the *CS* method. However, note that this comparison is not exactly fair, since *CS* is tested on data pre-selected using the other heuristics. A non-biased selection of data could lead to different *CS* performance. We plan to address this in our future work. Finally, we also think that the poor performance of the *CS* method for English-Greek is due at least in part to the performance of the machine translation system. For German the machine translation system is much better than for Greek, which is an under-resourced language, and this difference may well influence the results significantly.

## 5. Conclusion

In this work we described a framework for collecting comparable corpora from the web. To construct comparable corpora we start with news titles written in different languages, pair the titles and download only the corresponding article contents if the titles are comparable. To measure the comparability of two titles we investigated different heuristics. We showed that the best heuristics are *TS*, *HS* and *TLD* when used in combination.

Our technique is a promising and resource-light way of collecting comparable corpora likely to be of use for SMT, though further work needs to be done to confirm this. Thus, for future work we plan to extract fragments such as parallel sentences or phrases from our comparable corpora and investigate their impact on machine translation quality. Currently, we constrain titles to have at least five content words. By doing this we discard 50% of the Greek and 23% of the German articles and prevent them from being paired. To increase the recall of our method we aim to reduce these numbers by investigating further ideas for pairing.

**Acknowledgements** This work has been funded by the ACCURAT project, European Union Seventh Framework Programme (FP7/2007-2013), grant agreement no 248347 and by the European Union Seventh Framework Programme with grant agreement FP7-PEOPLE-2009-IIF-254562.

## 6. References

R. Barzilay and K. R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 50–57, Morristown, NJ, USA. Association for Computational Linguistics.

P.S. Braschler. 1998. Multilingual Information Retrieval Based on Document Alignment Techniques. In *Research*

[5] Titles which have less than five content words are not taken into consideration.

*and advanced technology for digital libraries: second European conference, ECDL'98, Heraklion, Crete, Cyprus, September 21-23, 1998: proceedings*, page 183. Springer Verlag.

C. Callison-Burch, P. Koehn, and M. Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 17–24, Morristown, NJ, USA. Association for Computational Linguistics.

H. Edmundson, P. 1969. New Methods in Automatic Extracting. *Journal of the Association for Computing Machinery*, 16:264–285.

D. Huang, L. Zhao, L. Li, and H. Yu. 2010. Mining large-scale comparable corpora from chinese-english news collections. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 472–480. Association for Computational Linguistics.

D. Kauchak and R. Barzilay. 2006. Paraphrasing for automatic evaluation. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 455–462, Morristown, NJ, USA. Association for Computational Linguistics.

T. Kumano, H. Tanaka, and T. Tokunaga. 2007. Extracting phrasal alignments from comparable corpora by using joint probability SMT model. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*, pages 95–103.

C. Lopez, V. Prince, and M Roche. 2011. Automatic titling of Articles Using Position and Statistical Information. *RANLP, 2011.*

Y. Marton, C. Callison-Burch, and P. Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 381–390. Association for Computational Linguistics.

S. Montalvo, R. Martinez, A. Casillas, and V. Fresno. 2006. Multilingual document clustering: an heuristic approach based on cognate named entities. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1145–1152. Association for Computational Linguistics.

D. S. Munteanu and D. Marcu. 2002. Processing comparable corpora with bilingual suffix trees. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 289–295, Morristown, NJ, USA. Association for Computational Linguistics.

D.S. Munteanu and D. Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

D. S. Munteanu and D. Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 81–88, Morristown, NJ, USA. Association for Computational Linguistics.

D.S. Munteanu, A. Fraser, and D. Marcu. 2004. Improved machine translation performance via parallel sentence extraction from comparable corpora. In *HLT-NAACL*, pages 265–272.

P. Nakov. 2008. Paraphrasing verbs for noun compound interpretation. In *Proc. of the Workshop on Multiword Expressions, LREC-2008.*

B. Pouliquen, R. Steinberger, C. Ignat, E. K
"asper, and I. Temnikova. 2004. Multilingual and cross-lingual news topic tracking. In *Proceedings of the 20th international conference on Computational Linguistics*, pages 959–es. Association for Computational Linguistics.

R. Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526. Association for Computational Linguistics.

P. Resnik. 1999. Mining the web for bilingual text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 527–534. Association for Computational Linguistics.

S. Sharoff, B. Babych, and A. Hartley. 2006. Using comparable corpora to solve problems difficult for human translators. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 739–746, Morristown, NJ, USA. Association for Computational Linguistics.

T. Talvensaari, A. Pirkola, K. Järvelin, M. Juhola, and J. Laurikkala. 2008. Focused web crawling in the acquisition of comparable corpora. *Information Retrieval*, 11(5):427–445.

S. Zhao, C. Niu, M. Zhou, T. Liu, and S. Li. 2008. Combining multiple resources to improve SMT-based paraphrasing model. In *Proceedings of ACL-08: HLT*, pages 1021–1029, Columbus, Ohio, June. Association for Computational Linguistics.