BUCC Shared Task: Cross-Language Document Similarity

Serge Sharoff
University of Leeds
Leeds, UK
s.sharoff@leeds.ac.uk

Pierre Zweigenbaum
LIMSI, CNRS
Orsay, France
pz@limsi.fr

Reinhard Rapp
University of Mainz
Mainz, Germany
reinhardrapp@gmx.de

Abstract

We summarise the organisation and results of the first shared task aimed at detecting the most similar texts in a large multilingual collection. The dataset of the shared was based on Wikipedia dumps with interlanguage links with further filtering to ensure comparability of the paired articles. The eleven system runs we received have been evaluated using the TREC evaluation metrics.

1 Task description

Parallel corpora of original texts with their translations provide the basis for multilingual NLP applications since the beginning of the 1990s. Relative scarcity of such resources led to greater attention to comparable (=less parallel) resources to mine information about possible translations. Many studies have been produced within the paradigm of comparable corpora, including publications in the BUCC workshop series since 2008.[1]

However, the community so far has not conducted an evaluation which compared different approaches for identifying more or less parallel documents in a large amount of multilingual data. Also, it is not clear how language-specific such approaches are. In this shared task we propose the first evaluation exercise, which is aimed at detecting the most similar texts in a large multilingual collection.

2 Data set

2.1 Description

The dataset is derived from static Wikipedia dumps of the main articles. A feature of Wikipedia is that it provides so-called inter-language links between many corresponding articles of different languages, i.e. between articles describing the same or corresponding headwords. These interlanguage links are provided by the authors of the articles, i.e. they are based on expert judgement. For the shared task we selected bilingual pairs of articles which fulfilled the following requirements:

1. The inter-language links between the articles had to be bidirectional, i.e. not only an article in Language$_1$ needs to be linked to the corresponding article in Language$_2$, but also vice versa. This ensured a page in one language is not linked only to a portion of a page in another one.
2. The size of the textual content of the two articles within a pair (i.e. their length measured as the number of characters) had to be similar (see Section 2.2 below).

Note that this selection procedure for the article pairs implies that an article pair selected for one language pair may or may not be selected for another language pair. All articles which satisfied the selection conditions have been considered for the evaluation run.

The data for each language pair has been split randomly into two sets:

Training set articles with information about the correct links for the respective language pairs provided to the participants;

Test set articles without the links.

The task is for each article in the test set to submit up to five ranked suggestions to its linked article, assuming that the gold standard contains its counterpart in another language. The submissions had to be in the tab-separated format as used in the submissions to the shared tasks of the Text Retrieval Conference (TREC[2]) with six fields:

[1] See http://comparable.limsi.fr/

[2] See http://trec.nist.gov/.

Proceedings of the Eighth Workshop on Building and Using Comparable Corpora, pages 74–78,
Beijing, China, July 30, 2015. ©2015 Association for Computational Linguistics

	Min.	1st Qu	Median	Mean	3rd Qu	Max.	Selected pairs
De	0.010	0.420	0.790	1.244	1.370	206.000	294990
Fr	0.000	0.370	0.740	1.194	1.260	255.800	229591
Ru	0.010	0.300	0.620	0.987	1.070	108.600	159810
Tr	0.000	0.140	0.350	0.616	0.760	46.730	32614
Zh	0.010	0.280	0.610	1.010	1.090	111.500	42944

Table 1: Ratios of lengths of aligned articles to English

```
    id1 X id2 Y score run.name
```

The X and Y fields are not used, but they are reserved by the TREC evaluation script (and it does not use them either). `id1` and `id2` are the respective article identifiers in a source language and in English. The *score* should reflect the similarity between `id1` and `id2`, the higher the closer. The participants were invited to submit up to five runs of their system with different parameters, as identified by a keyword in the last field.

The evaluation script and more information about the format have been made available in advance. [3]

The languages in the shared task were Chinese, French, German, Russian and Turkish. Pages in these languages needed to be linked to a page in English.

The choice of languages reflects variation in the available clues for linking the pages. The languages vary in:

- their writing systems (Latin, Cyrillic, logographic);
- tokenisation (clitics in French, compounds in German, no orthographic word boundaries in Chinese);
- their morphology (covering isolating, inflecting and agglutinative languages);

Even though the writing system issue is superficial, it shifts the clues for linking the articles. Thus, it requires more intelligent mapping between the languages. In the same writing system, many clues remain the same or nearly identical (*Paris, Frankfurt*), while in another set they have to adapt to the target language requirements: Париж ('Paris', transliterated as *Parizh* in Russian) or 巴黎 ('Paris', pinyin *Bali* in Chinese).

Morphology accounts for variation of forms for connection with the dictionaries. It is considerably larger in morphologically rich languages, such as Russian or Turkish. Therefore, mapping of word forms is likely to be more sparse.

2.2 Preparation

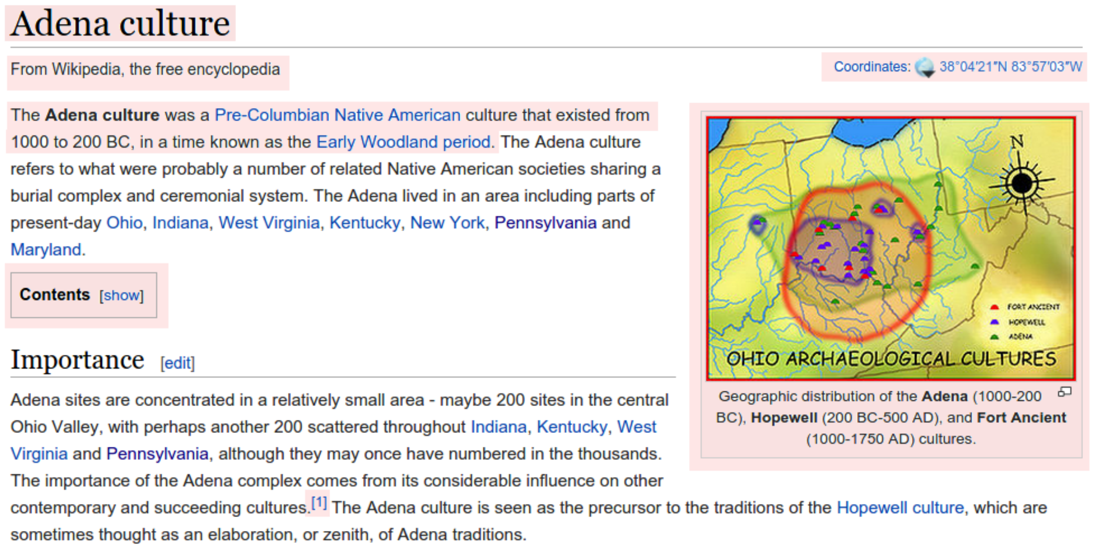
We started with the downloadable Wikipedia dumps,[4] which were cleaned to their text only contents by removing standard formatting codes, figures (with their captions), templates, tables and external links. Given that the first sentence in Wikipedia articles provide a concise summary of the article contents, the first sentence (defined as a sequence of characters to the first full stop) has been also removed to make the task more similar to detection of webpages in context unrelated to Wikipedia. Shaded areas in Figure 1 demonstrate the extent of cleaning.

We selected a subset of articles aligned to English. Table 1 lists the distribution of the length ratios of the respective articles to their English counterparts and the number of articles remaining after pruning their length. A small number of articles are much shorter than their English counterparts. Less frequently this happens in the opposite direction, and the length ratio is more than one (the median is always less than one). Usually articles which differ in their length are not good candidates for comparable corpora. We took only those within the inter-quartile range. This left us with 50% of article pairs in the original list, which are all reasonably comparable in their contents. Examples for each language bordering on the 1st quartile in ratio to English all show reasonable amount of text to be considered as comparable entries:

de	*Aaron Ramsey*
fr	*Adena culture*
ru	*Quantum mechanics*
tr	*Cyrano de Bergerac (play)*
zh	*Blood transfusion*

[3] See `http://trec.nist.gov/trec_eval/`

[4] Downloaded in November 2011.



Adena culture

From Wikipedia, the free encyclopedia

Coordinates: 38°04′21″N 83°57′03″W

The **Adena culture** was a Pre-Columbian Native American culture that existed from 1000 to 200 BC, in a time known as the Early Woodland period. The Adena culture refers to what were probably a number of related Native American societies sharing a burial complex and ceremonial system. The Adena lived in an area including parts of present-day Ohio, Indiana, West Virginia, Kentucky, New York, Pennsylvania and Maryland.

Contents [show]

Importance [edit]

Adena sites are concentrated in a relatively small area - maybe 200 sites in the central Ohio Valley, with perhaps another 200 scattered throughout Indiana, Kentucky, West Virginia and Pennsylvania, although they may once have numbered in the thousands. The importance of the Adena complex comes from its considerable influence on other contemporary and succeeding cultures.[1] The Adena culture is seen as the precursor to the traditions of the Hopewell culture, which are sometimes thought as an elaboration, or zenith, of Adena traditions.

Geographic distribution of the **Adena** (1000-200 BC), **Hopewell** (200 BC-500 AD), and **Fort Ancient** (1000-1750 AD) cultures.

Figure 1: Example of cleanup (shaded areas indicate removed text).

For example, the *Adena culture* article has been selected only for the French-English pair, since the articles in other languages are much shorter than the English one to be considered as reasonably comparable.

3 Evaluation

Evaluation has been done using standard TREC evaluation measures, modeling the task as the retrieval of a ranked list of links from a source page.

Extrinsic evaluation setups, for example, via terminology extraction, would possibly provide more interesting measures, but this would require a baseline system which works with all the languages in question.

3.1 Metrics

For each source page there exists exactly one correct linked page in the gold standard. Systems were required to return a ranked list of hypotheses in which the correct target page should be ranked as high as possible.

Several evaluation measures are relevant to this situation in the `trec_eval` program used in TREC evaluations. The *Success* measures correspond to commonly used measures when evaluating term translations in comparable corpora. We use them here to evaluate the proposed inter-language links between the articles. Success@1 determines the proportion of source articles for which the correct target article has been ranked in the top position; Success@5 determines the proportion of source articles for which the correct target article has been ranked among the top 5 positions. *Mean Reciprocal Rank* (MRR) is also a relevant measure: If the correct target article is ranked at position N, a score of $1/N$ is given to this source article. Then these scores are averaged over the set of source articles. These measures are respectively obtained by parameters `success.1`, `success.5`, and `recip_rank` in `trec_eval`.

4 Results

Overall, we have received eleven runs: one entry for Chinese (Table 2), three entries for French (Table 2), and seven for German (Table 3).

4.1 Methods used

The method used by the system CCNUNLP is described in (Li and Gaussier, 2013). In essence, it uses a bilingual dictionary for converting the word feature vectors between the languages and estimating their overlap. The other systems are discussed in details in the current proceedings (Morin et al., 2015; Zafarian et al., 2015). The LINA system (Morin et al., 2015) is based on matching hapax legomena, i.e., words occurring only once. In addition to using hapax legomena, the quality of linking in one language pair, e.g., French-English, is also assessed by using information available in pages in another language pair, e.g., German-English. The AUT system (Zafarian et al., 2015) uses the most complicated setup by combining several steps. First, documents in different languages are mapped into the same space using a

	French			Chinese
runid	ccnunlp	lina.p	lina.cl	ccnunlp
num_q	114423	114423	78529	21467
num_ret	572115	572111	143542	107335
num_rel	114423	114423	78529	21467
num_rel_ret	87367	42777	47561	18474
MRR	0.669	0.329	0.590	0.769
success@1	0.607	0.300	0.577	0.710
success@5	0.764	0.374	0.606	0.861

Table 2: Evaluation results for French and Chinese. `lina.p` corresponds to Pigeonhole, `lina.cl` to Cross-lingual in the authors' paper.

	German						
runid	lina.p	lina.cl	aut1	aut2	aut3	aut4	aut5
num_q	147220	92020	147515	147515	147515	147515	147515
num_ret	736100	166051	147516	147516	147516	147516	147516
num_rel	147220	92020	147515	147515	147515	147515	147515
num_rel_ret	52223	58828	6870	2703	2029	1371	890
MRR	0.290	0.622	0.047	0.018	0.014	0.009	0.006
success@1	0.249	0.607	0.047	0.018	0.014	0.009	0.006
success@5	0.355	0.639	0.047	0.018	0.014	0.009	0.006

Table 3: Evaluation results for German

feature transformation matrix. This helps in selecting a relatively small subset of pages to detect possible links. Second, document similarity is assessed using three pipelines, namely, a polylingual topic model, a named entities detection tool and a word feature mapping procedure using MT.

4.2 Comparison of results

Since AUT submitted exactly one target article for each source article, its MRR, success@1 and success@5 measures are identical.

For each run, success@1 is the strictest measure, hence provides the lowest score, because it can only obtain points if the top ranked article is the correct one. Mean reciprocal rank (MRR) yields the same score when the top ranked article is correct, but also scores decreasing fractions of one when the correct article is found anywhere in the ranking: this results in a higher average score than success@1. Finally, success@5 also takes into account articles beyond the first, but only until the fifth; if the correct article is present in this range, the full score of one is assigned to the article; otherwise no point is obtained. Therefore a system which generally ranks correct articles beyond the fifth position will have a lower success@5 than its MRR; but a system which ranks correct articles before the sixth position often enough will have a higher success@5 than MRR. This is the case of all systems except aut, which only returned one target article per source article.

The tables show that the rankings obtained by the three measures, MRR, success@1 and success@5 are the same in all cases, i.e., rank correlation of the results is always 1. This suggests that system results ranked the correct article in the top 5 often enough.

4.3 Comparison of methods

The best results were obtained on Chinese with a succes@1 of 0.710 and a success@5 of 0.861. This is a very good performance, but also reveals that the problem is not solved.

Although the number of different runs is not sufficient to draw general conclusions, we can compare the same methods across different language pairs and different methods on the same language pairs.

CCNUNLP obtained better results on Chinese than on French, probably because of the quality of the underlying dictionaries. LINA.CL worked better on German than for French, while the reverse was true for LINA.P. After the evaluation run, it transpired that the submissions of AUT had a data processing bug.

Overall, the CCNUNLP method obtained the best results on Chinese and French, followed by the LINA.CL method (second best on French, and best on German).

4.4 Discussion

The results are encouraging. Success@1 rates reach 0.71 for Chinese and 0.61 for French and German. However, this level of accuracy is still far from reliable identification of comparable pages. Given a small number of participating systems and an uneven coverage of the language pairs involved it is difficult to make predictions about which methods are more or less successful. A dictionary-based method (CCNUNLP) is slightly ahead of a method based on hapax legomena (LINA.*). A multi-stage method like the one used by AUT is promising, but its complexity makes it prone to errors.

Another question concerns the evaluation scenario. The shared task has been evaluated by using gold standard data in intrinsic evaluation. Given that the purpose of collecting comparable corpora is to provide more data for terminology extraction or Machine Translation, we need to evaluate text collections by referring to their successful use in such tasks. The limitation in using extrinsic evaluation is the lack of gold-standard methods and resources.

In the next shared task we plan to address this issue by specifically targeting either terminology extraction or MT development methods by using comparable corpora. This shared task will use the resources we developed for the current one.

4.5 Conclusions

In addition to obtaining an estimate of the quality of various methods for measuring comparability, the major outcomes of the evaluation exercise concerns the available standardised dataset which is split into the training and testing parts. We encourage our readers to develop better systems and to test them on our data. The dataset is available from:

`http://corpus.leeds.ac.uk/serge/BUCC/`

We intend to keep the data on the web for many years as a benchmark for measuring comparability on the text level.

References

Li, B. and Gaussier, E. (2013). Exploiting comparable corpora for lexicon extraction: Measuring and improving corpus quality. In Sharoff, S., Rapp, R., Zweigenbaum, P., and Fung, P., editors, *Building and Using Comparable Corpora*, pages 131–149. Springer-Verlag.

Morin, E., Hazem, A., Boudin, F., and Loginova-Clouet, E. (2015). Lina: Identifying comparable documents from wikipedia. In *Proc. Workshop on Building and Using Comparable Corpora at ACL 2015.*

Zafarian, A., Agha Sadeghi, A. P., Azadi, F., Ghiasifard, S., Ali Panahloo, Z., bakhshaei, S., and Mohammadzadeh Ziabary, S. M. (2015). Aut document alignment framework for bucc workshop shared task. In *Proc. Workshop on Building and Using Comparable Corpora at ACL 2015.*