Comparing Multilingual Comparable Articles Based On Opinions

Motaz Saad **David Langlois** **Kamel Smaïli**
Speech Group, LORIA
INRIA, Villers-lès-Nancy, F-54600, France
Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
CNRS, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
{firstName.lastName}@loria.fr

Abstract

Multilingual sentiment analysis attracts increased attention as the massive growth of multilingual web contents. This conducts to study opinions across different languages by comparing the underlying messages written by different people having different opinions. In this paper, we propose Sentiment based Comparability Measures (SCM) to compare opinions in multilingual comparable articles without translating source/target into the same language. This will allow media trackers (journalists) to automatically detect public opinion split across huge multilingual web contents. To develop SCM, we need either to get or to build parallel sentiment corpora. Because this kind of corpora are not available, we decided to build them. For that, we propose a new method to automatically label parallel corpora with sentiment classes. Then we use the extracted parallel sentiment corpora to develop multilingual sentiment analysis system. Experimental results show that, the proposed measure can capture differences in terms of opinions. The results also show that comparable articles variate in their objectivity and positivity.

1 Introduction

We can distinguish two kinds of sentiments analysis depending on monolingual or multilingual articles.

In the following, as in (Pang and Lee, 2008), the terms Sentiment Analysis (SA) and Opinion Mining (OM) are used as synonyms. Mining opinions is to identify the subjectivity and/or the polarity of a given text at article or sentence level. Subjectivity identification is to classify the text into subjective or objective, while polarity identification is to classify the text into negative or positive.

Popular methods for monolingual sentiment analysis are based on lexicon and corpus. Lexicon based methods use string matching techniques between texts and annotated lexicons. The most common sentiment lexicons for English language are WordNet-Affect (Valitutti, 2004) and SentiWordNet (Esuli and Sebastiani, 2006), which are extensions of WordNet. Additionally, SenticNet (Cambria et al., 2010) is a knowledge-base extension of aforementioned lexicons. On the other hand, corpus based approach is popular for sentiment analysis (Pang and Lee, 2008). It uses corpora and machine learning algorithms to build sentiment classification systems. For example, Pang *et al.* used polarity (Pang et al., 2002) and subjectivity (Pang and Lee, 2004) English corpora to train machine learning algorithms to build sentiment classifiers. These resources have been adapted to other languages by many researchers as we will see in the following.

Multilingual sentiments analysis becomes a reality because of the massive growth of multilingual web contents. In this case, sentiment analysis identifies sentiments across multiple languages instead of one language. This can be done by creating sentiment resources for new languages by translating existing English resources (lexicons/corpora) into the target language, or by translating target text into English, then pass the translated text to English models for sentiment analysis (Rushdi-Saleh et al., 2011; Bautin et al., 2008; Denecke, 2008; Ghorbel, 2012). However, (Brooke et al., 2009) reported that creating new resources to build sentiment models from scratch works better than using the approach based on machine translation.

As we see in the previous discussion, works on multilingual sentiment analysis just try to identify sentiments across multiple languages. How-

Proceedings of the 6th Workshop on Building and Using Comparable Corpora, pages 105–111,
Sofia, Bulgaria, August 8, 2013.

ever, it is worthy to compare opinions about a given topic in several languages, not just to identify these opinions. If people from different cultures wrote an article about political/societal topics, they may judge these topics differently according to their cultures. In fact, detecting disagreement of opinions in multiple languages is a promising research area. So, our goal is to enable media trackers (journalists) to automatically detect the split of public opinions about a given topic across multiple languages. To the best of our knowledge, there are no work in the literature that serve our goal, therefore, we propose to develop automatic measures that compare opinions in multilingual comparable articles. These comparability measures will be the core of our goal which is building multilingual automatic journalist review system.

For that, we propose a Sentiment based Comparability Measures (SCM) which identify sentiments, score them and compare them across multilingual documents. Therefore, we need to identify and score sentiments in multiple languages. Namely, SCM needs a multilingual sentiment analysis system to identify and score sentiments. To build this system, we need parallel sentiment corpora from different topics. Unfortunately, we do not have such corpora, we only have English sentiment corpus. So, we propose in Section 2 a new method to build parallel sentiment corpora. We start from English sentiment corpora (movie reviews domain), then use it to build sentiment classifier for English language and then label a new parallel English/target corpora which is different from the movie one. In section 3, we use the obtained parallel sentiment corpora to build a multilingual sentiment analysis system which is used to develop SCM, then we use SCM to compare multilingual comparable articles in terms of opinions. The advantage of this idea is that we do not need to translate corpora/lexicons to analyse multilingual text.

The rest of this article is organized as follows, Section 2 describes our method to build parallel sentiment corpora, Section 3 presents our proposed sentiment based comparability measures (SCM) and experimental results conducted on corpora. Finally, we state the conclusions.

2 Sentiment Corpora Extraction

As we introduced earlier, we need parallel corpora to build the sentiment comparability measure. Therefore, we present in this section a method to annotate parallel corpora with sentiment labels. This method can be applied on any English/target language pairs. In this work, we label English/Arabic parallel sentences. The idea is to use an English sentiment classifier to label each English sentence in the new parallel corpora, then we can assign the same label to the target (Arabic) sentence, because sentences are parallel and convey the same opinions.

The widely used approach to build a classifier is to build a Naive Bayes model using n-grams linguistic features (Pang et al., 2002; Dave et al., 2003; Pang and Lee, 2004; Kim and Hovy, 2004; Cui et al., 2006; Tan et al., 2009). So, we use this method on bigrams extracted from English sentiment corpora of movie reviews. These corpora are manually labelled with subjectivity and polarity labels. Each review in the collection is represented as a vector composed of bigram occurrences. Then, each vector is feed to Naive Bayes classifier with corresponding class label for training. Naive Bayes classifies the vector to the highest probable class. Our objective in this paper is to compare opinions, this is why we used this traditional method for building the sentiment classifier.

The parallel corpora, that we annotate, cover variant topics (newspapers, UN resolutions, and transcribed talks), and are available in many languages. The newspapers are collection of parallel articles from AFP, ANN, ASB, and provided by LDC[1]. UN corpora[2] is a collection of United Nations General Assembly Resolutions. Transcribed talks are collection of multilingual transcriptions from TED provided by WIT3[3].

Figure 1 illustrates our method and Table 1 describes corpora denoted in the figure. The mentioned corpora are: *senti-corp*, *parallel*, and *new-senti-corp*. *senti-corp* represents the monolingual (English) manually labelled, *parallel* represents parallel corpora in variant topics, and *new-senti-corp* represents the extracted corpora. Corpora sizes are presented in Tables 2 and 3. Table 2 presents the number of reviews of *senti-corp* with

[1]LDC - Linguistic Data Consortium: ldc.upenn.edu

[2]Corpora of the United Nations: uncorpora.org

[3]WIT3 Web Inventory of Transcribed and Translated Talks wit3.fbk.eu

respect to sentiment classes, and Table 3 presents the number of sentences of parallel corpora.

Table 1: Corpora description

Corpora	Description
senti-corp	Monolingual manually labelled sentiment corpus (polarity or subjectivity)
senti-corp-p1	Part 1 of senti-corp (90%): used to build classification models which are used for labelling task
senti-corp-p2	Part 2 of senti-corp (10%): This is the (test corpus) which is used to test the extracted corpora
parallel	Multilingual parallel corpora
parallel-p1	Part 1 of the parallel corpora (90%): to be labelled automatically
parallel-p2	Part 2 of the parallel corpora (10%): to be used to evaluate SCM
new-senti-corp	Multilingual automatically labelled sentiment corpus

Table 2: *Senti-cor*p size (number of reviews)

Class	senti-corp-p1	senti-corp-p2
subjective	4500	500
objective	4500	500
negative	900	100
positive	900	100

Table 3: *Parallel* Corpora size

Corpus	# of sentences
parallel-p1	364K
parallel-p2	40K

The following steps describe the method we propose:

1. Split *senti-corp* into two parts: *senti-corp-p1* is 90%, and *senti-corp-p2* is 10%.

2. Use *senti-corp-p1* to train a Naive Bayes classifier to build a monolingual sentiment model.

3. Split the parallel corpora into two parts: *parallel-p1* is 90%, and *parallel-p2* is 10%.

4. Using the sentiment classification model obtained in step 2, classify and label English sentences of *parallel-p1* and assign the same sentiment class to the corresponding Arabic sentences.

5. Refine and filter sentences which are labelled in step 4. The filtering process keeps only sentences that have high sentiment score. Then, we obtain *new-senti-corp* which is Arabic/English parallel sentiment labelled corpora in different domains.

6. Use the English part of *new-senti-corp* which is obtained in step 5 to train a Naive Bayes classifier.

7. Evaluate the classifier built in step 6 on *senti-corp-p2*. If the classification accuracy is accepted, then continues, otherwise, try other corpora and/or models.

This method is independent of the the sentiment class labels. So, it can be applied for subjectivity or polarity corpus.

Tables 4 and 5 present the experimental results of steps 4 and 5 of the Figure 1. Table 4 shows the statistical information of sentiment scores of the labelled corpora, where Rate is the class label distribution (percentage) with respect to the whole dataset. μ, σ, Min, and Max are the mean, standard deviation, minimum, and maximum values of sentiment scores respectively. For subjectivity labels, 54% and 46% of sentences are labelled as subjective and objective respectively. For polarity labels, 58% and 42% of sentences are labelled as negative and positive respectively. Table 5 presents the frequency table of intervals of sentiment scores of the labelled sentences. We can see from Table 5 that most of sentences have high sentiment scores (from 0.9 to 1.0). To extract high quality labelled sentences, we keep only sentences with score greater than 0.8.

In order to evaluate the quality of the extracted corpora (step 7 in Figure 1), we need first to build a sentiment classifier based on this corpora and then evaluate the accuracy of this classifier. The detail of this process is given bellow:

1. Train a Naive Bayes classifier on the parallel sentiment corpora *new-senti-corp*.

2. Test the obtained classifiers on the manually labelled corpus *senti-corp-p2*.

Figure 1: Approach for parallel sentiment corpora extraction and evaluation

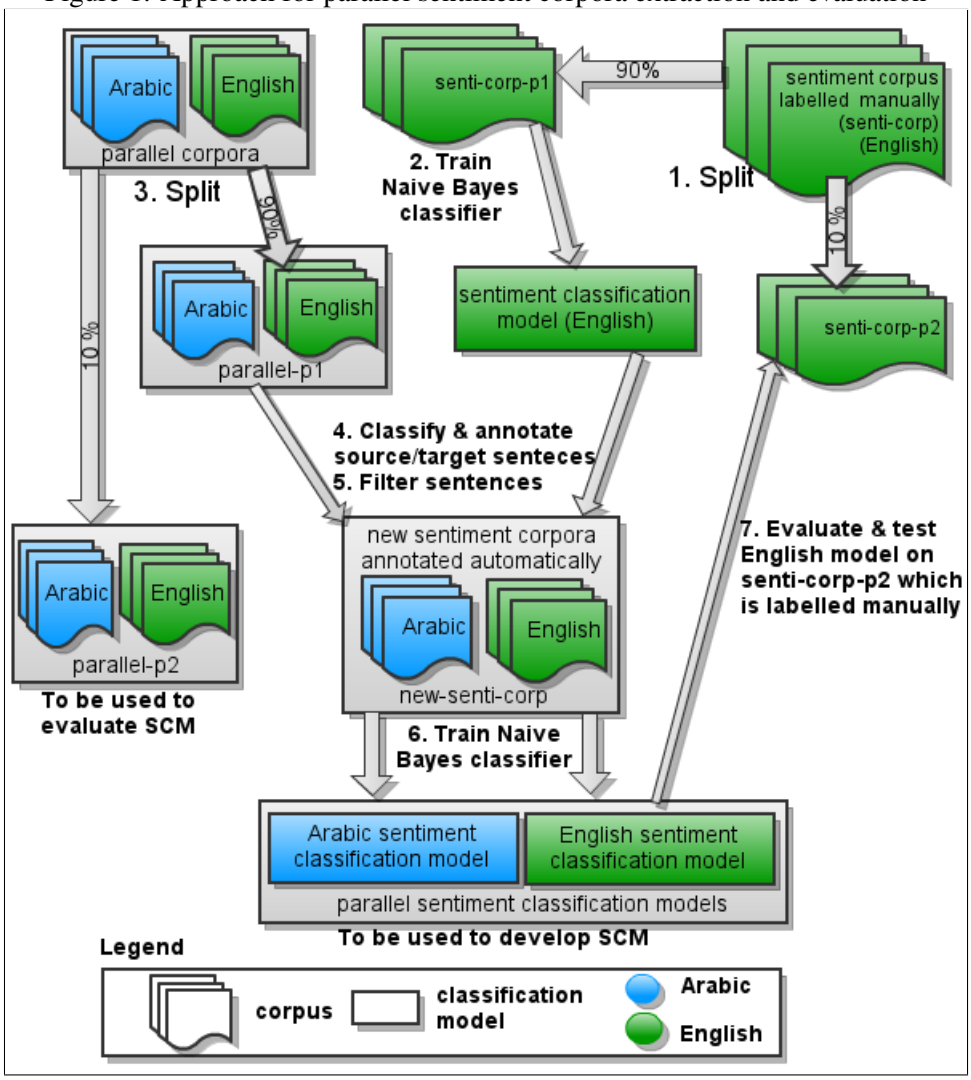

Table 4: Sentiment classes statistics for labelled sentences scores of *parallel-p1* corpora

Label	**Count**	**Rate**	μ	σ	**Min**	**Max**
subjective	231,180	54%	0.93	0.11	0.60	1.00
objective	197,981	46%	0.93	0.11	0.60	1.00
negative	219,070	58%	0.84	0.12	0.60	0.99
positive	159,396	42%	0.83	0.12	0.60	1.0

Table 5: Frequency table of sentiment scores intervals of labelled sentences of *parallel-p1* corpora

Label	**[0.6,0.7)**	**[0.7,0.8)**	**[0.8,0.9)**	**[0.9,1]**
subjective	6.1%	9.0%	11.9%	73.0%
objective	6.8%	8.1%	10.8%	74.3%
negative	17.7%	18.0%	21.6%	42.7%
positive	20.4%	20.8%	21.7%	37.2%

In the following, *senti-corp-p2* is the test corpus. The evaluation is presented in Table 6. The metrics include classification accuracy, and F-measures. F-neg, F-pos, F-sub, and F-obj are the F-measures for negative, positive, subjective, and objective classes respectively. For subjectiv-

Table 6: Evaluation of extracted corpus (step 7)

Subjectivity		**Polarity**	
Accuracy	0.765	Accuracy	0.720
F-sub	0.717	F-neg	0.754
F-obj	0.799	F-pos	0.674

ity test, the classifier achieved 76.5% of accuracy and an average of 75.8% of f-measure. For polarity test, the classifier leads to 72% of accuracy and an average of 71% of F-measure.

We wanted to compare these results with others works in sentiment classification, but unfortunately the used corpora are not the same. Anyway, these results are only indicative for us, because our objective is not to propose a new method for automatic sentiment classification, but to build a sentiment based comparability measure.

Now, we obtained English/Arabic parallel sentiment corpora in multiple topics. We use these corpora to develop sentiment based comparability measures that will be described in the next section.

Notice that at the beginning the only available sentiment corpus was a collection of movie reviews in English language, with the proposed method, we got multilingual sentiment corpora of different topics. Furthermore, using this method, one can obtain sentiment corpus for under-resourced languages. The advantage of the parallel corpora is to build sentiment classifiers that can be used to develop sentiment based comparability measures.

3 Sentiment Based Comparability Measures

As we stated in the introduction, there are no work in the literature that serve our goal, which is to compare multilingual articles in terms opinions. Therefore, we propose to develop automatic measures that compare opinions in multilingual comparable articles.

In the previous section, we built a parallel sentiment corpora where both source and its corresponding sentence have the same sentiment label. In this section, we compare multilingual comparable articles in terms of sentiments. Obviously, in this case we do not have the same sentiment labels since articles are comparable and not parallel. So, we develop Sentiment based Comparability Measures (SCM) which measure the differences of opinions in multilingual corpora. For that, we use the achieved parallel sentiment corpora *new-senti-corp* to build multilingual sentiment analysis systems, using the same method as in Section 2.

The idea is to identify and score sentiments in the source and target comparable articles and provide these information to SCM to compare their opinions. In the following, we describe how to compute SCM for comparable articles based on average score of all sentences.

We use formula 1 which is derived from Naive Bayes to compute opinion score and assign the corresponding label:

$$classify(S) = \underset{c}{argmax} P(c) \prod_{k=1}^{n} P(f_k|c) \quad (1)$$

where S is a sentence, f_k are the features of S, $c \in \{o, \bar{o}\}$ for subjectivity and $c \in \{p, \bar{p}\}$ for polarity, where o is objective, $\bar{o}$ is subjective, p is positive, $\bar{p}$ is negative.

An article may contain some sentences belonging to the subjective class, and others belonging to the objective class (idem for positive and negative). So, for a given pair of comparable articles, SCM has three parameters d_x, d_y, c, where d_x, d_y are the source and the target articles respectively, and c is the class label. This score is calculated as follows:

$$SCM(d_x, d_y, c) = \left| \frac{\sum\limits_{C(S_x)=c} P(S_x|c)}{N_x} - \frac{\sum\limits_{C(S_y)=c} P(S_y|c)}{N_y} \right| \quad (2)$$

Where $S_x \in d_x$, $S_y \in d_y$, and $\sum\limits_{C(S_x)=c} P(S_x|c)$ and $\sum\limits_{C(S_y)=c} P(S_y|c)$ are the sum of probabilities for all source and target sentences respectively that belong to class c. N_x and N_y are the number of source and target sentences respectively that belong to the class c. Formally speaking, for a given pair of documents d_x, d_y, we have four measures: $SCM(d_x, d_y, o)$, $SCM(d_x, d_y, \bar{o})$ for subjectivity, and $SCM(d_x, d_y, p)$, $SCM(d_x, d_y, \bar{p})$ for polarity.

In our experiments, we calculate SCM for pair of articles in parallel and comparable corpora. Calculating SCM for parallel corpora could be very surprising, but we did it in order to show that for this kind of corpora, the proposed measure should be better than the one achieved for comparable corpora.

Table 7: Comparable corpora information

	AFEWC		eNews	
	English	Arabic	English	Arabic
Articles	40290	40290	34442	34442
Sentences	4.8M	1.2M	744K	622K
Average #sentences/article	119	30	21	17
Average #words/article	2266	548	198	161
Words	91.3M	22M	6.8M	5.5M
Vocabulary	2.8M	1.5M	232K	373K

Table 8: Average Sentiment Based Comparability Measures (SCM)

Corpora		$SCM(d_x, d_y, \bar{o})$	$SCM(d_x, d_y, o)$	$SCM(d_x, d_y, \bar{p})$	$SCM(d_x, d_y, p)$
parallel-p2	AFP	0.02	0.02	0.1	0.12
	ANN	0.05	0.06	0.1	0.1
	ASB	0.07	0.1	0.12	0.14
	TED	0.06	0.06	0.08	0.07
	UN	0.05	0.02	0.07	0.08
Comparable	ENews	0.07	0.15	0.11	0.15
	AFEWC	0.11	0.19	0.11	0.16

The comparable corpora that we use for our experiments are AFEWC and eNews which were collected and aligned at article level (Saad et al., 2013). Each pair of comparable articles is related to the same topic. AFEWC corpus is collected from Wikipedia and eNews is collected from Euronews website. Table 7 presents the number of articles, sentences, average sentences per article, average words per article, words, and vocabulary of these corpora.

Table 8 presents the experimental results of SCM computed using formula 2. SCM is computed for the source and target articles for parallel corpora *parallel-p2* and comparable corpora (AFEWC and eNews). We note that SCM for AFP, ANN, ASB, TED, and UN corpora are small because they are parallel. This shows that the proposed measure is well adapted to capture the similarity between parallel articles. Indeed, they have the same sentiments. On the other hand, SCM become larger for comparable corpora, because the concerned articles do not necessary have the same sentiments. The only exception to what have been claimed is that the subjectivity SCM for eNews comparable corpora is similar to the one of ASB which is parallel corpora. In contrast, the objectivity SCM is larger (0.15) for eNews, that means pair of articles in eNews corpora have similar subjective but different objective sentiments. In other words, source and target are considered similar in terms of subjectivity but different in terms of objectivity (idem for negative and positive). Consequently, comparable articles do not necessary have the same opinions. Additionally, we note that the SCM for AFEWC corpora are the largest in comparison to the others, this is maybe because Wikipedia has been written by many different contributors from different cultures.

4 Conclusions

We presented a new method for comparing multilingual sentiments through comparable articles without the need of translating source/target articles into the same language. Our results showed that it is possible now for media trackers to automatically detect difference in public opinions across huge multilingual web contents. The results showed that the comparable articles variate in their objectivity and positivity. To develop our system, we required parallel sentiment corpora. So, we presented in this paper an original method to build parallel sentiment corpora. We started from an English movie corpus annotated in terms of sentiments, we trained NB classier to classify an English text concerning topics different from movie, and then we deduced the sentiment labels of the the corresponding target parallel text by assigning the same labels. This method is interest-

ing because it allows us to produce several parallel sentiment corpora concerning different topics. We built SCM using these parallel sentiment corpora, then, SCM identifies sentiments, scores them and compares them across multilingual documents. In the future works, we will elaborate our journalist review system by developing a multilingual comparability measure that can handle semantics and integrate it with the sentiment based measure.

References

M. Bautin, L. Vijayarenu, and S. Skiena. 2008. International sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

J. Brooke, M. Tofiloski, and M. Taboada. 2009. Cross-linguistic sentiment analysis: From english to spanish. In *International Conference RANLP*, pages 50–54.

E. Cambria, R. Speer, C. Havasi, and A. Hussain. 2010. Senticnet: A publicly available semantic resource for opinion mining. *Artificial Intelligence*, pages 14–18.

H. Cui, V. Mittal, and M. Datar. 2006. Comparative experiments on sentiment classification for online product reviews. In *proceedings of the 21st national conference on Artificial intelligence - Volume 2*, AAAI'06, pages 1265–1270. AAAI Press.

K. Dave, S. Lawrence, and D. M. Pennock. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 519–528, New York, NY, USA. ACM.

K. Denecke. 2008. Using sentiwordnet for multilingual sentiment analysis. In *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*, pages 507–512.

A. Esuli and F. Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation*, pages 417–422.

H. Ghorbel. 2012. Experiments in cross-lingual sentiment analysis in discussion forums. In K. Aberer, A. Flache, W. Jager, L. Liu, J. Tang, and C. Guret, editors, *Social Informatics*, volume 7710 of *Lecture Notes in Computer Science*, pages 138–151. Springer Berlin Heidelberg.

S.-M. Kim and E. Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

B. Pang and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.

B. Pang and L. Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

M. Rushdi-Saleh, M. T. Martín-Valdivia, L. A. Ureña López, and J. M. Perea-Ortega. 2011. Bilingual experiments with an arabic-english corpus for opinion mining. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 740–745, Hissar, Bulgaria, September. RANLP 2011 Organising Committee.

M. Saad, D. Langlois, and K. Smaïli. 2013. Extracting comparable articles from wikipedia and measuring their comparabilities. In *V International Conference on Corpus Linguistics*. University of Alicante, Spain.

S. Tan, X. Cheng, Y. Wang, and H. Xu. 2009. Adapting naive bayes to domain adaptation for sentiment analysis. In *Advances in Information Retrieval*, pages 337–349. Springer.

R. Valitutti. 2004. Wordnet-affect: an affective extension of wordnet. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086.

Mining for Domain-specific Parallel Text from Wikipedia

Magdalena Plamadă, Martin Volk
Institute of Computational Linguistics, University of Zurich
Binzmühlestrasse 14, 8050 Zurich
{plamada, volk}@cl.uzh.ch

Abstract

Previous attempts in extracting parallel data from Wikipedia were restricted by the monotonicity constraint of the alignment algorithm used for matching possible candidates. This paper proposes a method for exploiting Wikipedia articles without worrying about the position of the sentences in the text. The algorithm ranks the candidate sentence pairs by means of a customized metric, which combines different similarity criteria. Moreover, we limit the search space to a specific topical domain, since our final goal is to use the extracted data in a domain-specific Statistical Machine Translation (SMT) setting. The precision estimates show that the extracted sentence pairs are clearly semantically equivalent. The SMT experiments, however, show that the extracted data is not refined enough to improve a strong in-domain SMT system. Nevertheless, it is good enough to boost the performance of an out-of-domain system trained on sizable amounts of data.

1 Introduction

A high-quality Statistical Machine Translation (SMT) system can only be built with large quantities of parallel texts. Moreover, systems specialized in specific domains require in-domain training data. A well-known problem of SMT systems is that existing parallel corpora cover a small percentage of the possible language pairs and very few domains. We therefore need a language-independent approach for discovering parallel sentences in the available multilingual resources.

This idea was explored intensively in the last decade with different text sources, generically called comparable corpora, such as news feeds, encyclopedias or even the entire Web. The first approaches focused merely on news corpora and were either based on IBM alignment models (Zhao and Vogel, 2002; Fung and Cheung, 2004) or employing machine learning techniques (Munteanu and Marcu, 2005; Abdul Rauf and Schwenk, 2011).

The multilingual Wikipedia is another source of comparable texts, not yet thoroughly explored. Adafre and de Rijke (2006) describe two methods for identifying parallel sentences across it based on monolingual sentence similarity (MT and respectively, lexicon based). Fung et al. (2010) approach the problem by combining recall- and precision-oriented methods for sentence alignment, such as the DK-vec algorithm or algorithms based on cosine similarities. Both approaches have achieved good results in terms of precision and recall.

However, we are interested in real application scenarios, such as SMT systems. The following approaches report significant performance improvements when using the extracted data as training material for SMT: Smith et al. (2010) use a maximum entropy-based classifier with various feature functions (e.g. alignment coverage, word fertility, translation probability, distortion). Ştefănescu et al. (2012) propose an algorithm based on cross-lingual information retrieval, which also considers similarity features equivalent to the ones mentioned in the previous paper.

The presented approaches extract general purpose sentences, but we are interested in a specific topical domain. We have previously tackled the problem (Plamada and Volk, 2012) and encountered two major bottlenecks: the alignment algorithm for matching possible candidates and the similarity metric used to compare them. To our knowledge, existing sentence alignment algorithms (including the one we have employed in the first place) have a monotonic order constraint, meaning that crossing alignments are not

Proceedings of the 6th Workshop on Building and Using Comparable Corpora, pages 112–120,
Sofia, Bulgaria, August 8, 2013.