Building and Modelling Multilingual Subjective Corpora

Motaz Saad, David Langlois, Kamel Smaïli

Building and Modelling Multilingual Subjective Corpora

Motaz Saad, David Langlois, Kamel Smaïli

SMa^rT 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
{first.surname}@loria.fr

Abstract

Building multilingual opinionated models requires multilingual corpora annotated with opinion labels. Unfortunately, such kind of corpora are rare. We consider opinions in this work as subjective or objective. In this paper, we introduce an annotation method that can be reliably transferred across topic domains and across languages. The method starts by building a classifier that annotates sentences into subjective/objective label using a training data from "movie reviews" domain which is in English language. The annotation can be transferred to another language by classifying English sentences in parallel corpora and transferring the same annotation to the same sentences of the other language. We also shed the light on the link between opinion mining and statistical language modelling, and how such corpora are useful for domain specific language modelling. We show the distinction between subjective and objective sentences which tends to be stable across domains and languages. Our experiments show that language models trained on objective (respectively subjective) corpus lead to better perplexities on objective (respectively subjective) test.

Keywords: subjectivity analysis, cross-lingual annotation, language modelling

1. Introduction

Opinion mining or sentiment analysis determines the position of a writer with respect to a topic (Pang and Lee, 2008). One of opinion mining tasks is subjectivity identification, which is to classify a given text into subjective or objective (Pang and Lee, 2008).

A subjective or opinionated text conveys the opinions of the writer, while an objective text represents facts. For example, the following statement: "I think it is cold" is subjective because it conveys the opinion of a specific person. This opinion can be different from a person to another (a person lives in Mediterranean climate vs. a person lives in tundra climate). On the other hand, the following statement: "the temperature is (-15)" is objective because it represents a fact. However, for controversial statements, humans may disagree between themselves about the subjectivity of these statements. This is why automatic subjective analysis is a challenging task.

According to the opinion definition above, subjective and objective texts are very different in terms of writing style. Consequently, language models built from subjective texts may differ from ones built from objective texts. Therefore, opinion language models built from opinion corpora can be useful for certain domains and applications because they are better and more precise than non-domain specific language models. The distinction between multilingual subjective/objective text is investigated in the experiments of this work.

One of the wide applications of language modelling is text/speech machine translation, which require multilingual corpora. Since subjective/objective texts are distinct as mentioned earlier, then building multilingual subjective/objective corpora is quite useful for building machine translation systems in this domain.

Building opinionated language models requires corpus of both modalities (subjective and objective). So, Section 2. describes an automatic method (Saad et al., 2013a) to annotate multilingual parallel corpus with opinion labels. The parallel corpus represents different domain topics (United Nations resolutions, newspapers, and talks). This work focus on English/Arabic parallel corpus. Annotating this languages pair can be useful because such corpus is not available. Moreover, cross-lingual annotation can be used in information extraction in low-resourced languages.

Then, Section 3. shows how statistical language models trained on opinion corpora obtain better performance on test corpora. In addition, the link between subjectivity annotation and language modelling is studied in this section.

2. Subjective Corpora Annotation

This section describes the method of automatic subjective annotation of parallel corpora with opinion labels. The method requires a pre-annotated monolingual corpus to automatically annotate a parallel corpus. Before describing the method, these corpora are described here.

2.1. Corpora Description

The pre-annotated monolingual corpus is a collection of movie reviews. It is in English language and composed of 5,000 subjective and 5,000 objective sentences. The corpus was collected and annotated by (Pang and Lee, 2004). Subjective reviews were obtained from Rotten Tomatoes website `www.rottentomatoes.com`, while Objective reviews were obtained from IMDb plot summaries `www.imdb.com`. A Sample of subjective/objective sentences are presented in Tables 1 and 2.

Regarding parallel corpus, it is described in Table 3, where $|S|$ is the number of sentences, $|W|$ is the number of words, and $|V|$ is the vocabulary size. The corpora are collected from different sources and represent different genres of text.

pretty much sucks, but has a funny moment or two.
smart and alert, thirteen conversations about one thing is a small gem works both as an engaging drama and an incisive look at the difficulties facing native Americans.
even a hardened voyeur would require the patience of job to get through this interminable, shapeless documentary about the swinging subculture.
when perry fists a bull at the Moore farm, it's only a matter of time before he gets the upper hand in matters of the heart.

Table 1: A sample of subjective reviews of movie corpus

spurning her mother's insistence that she get on with her life, Mary is thrown out of the house, rejected by Joe, and expelled from school as she grows larger with child.
Amitabh can't believe the board of directors and his mind is filled with revenge and what better revenge than robbing the bank himself, ironic as it may sound.
the movie begins in the past where a young boy named Sam attempts to save Celebi from a hunter.

Table 2: A sample of objective reviews of movie corpus

AFP[1], ANN[2], and ASB[3] news corpora were provided by (Ma and Zakhary, 2009). Medar news corpus was provided by `www.medar.info`. NIST[4] news corpus was provided by (NIST, 2010). UN corpus is resolutions of United Nation and was provided by (Rafalovitch and Dale, 2009). TED[5] talks corpus was provided by (Cettolo et al., 2012). Tatoeba[6] parallel sentences was provided by (Tiedemann, 2012).

Corpus	$\|S\|$	$\|W\|$		$\|V\|$	
		English	Arabic	English	Arabic
AFP	4K	140K	114K	17K	25K
ANN	10K	387K	288K	39K	63K
ASB	4K	187K	139K	21K	34K
TED	88K	1.9M	1.6M	88K	182K
UN	61K	2.8M	2.4M	42K	77K
Medar	13K	398K	382K	43K	71K
NIST	2K	85K	64K	15K	22K
Tatoeba	1K	17K	13K	4K	6K
Total	183K	5.9M	5M	269K	480K

Table 3: Parallel Corpora

2.2. Automatic Annotation

The method (Saad et al., 2013a) automatically annotates English/Arabic parallel corpus with the help of English corpus which is composed of movie reviews and pre-annotated with opinion labels.

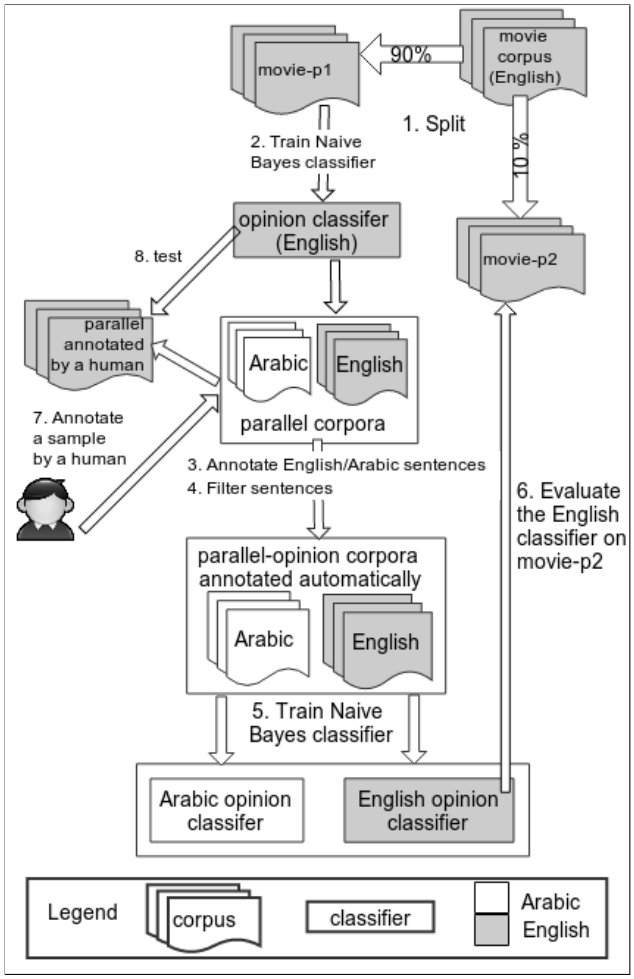

Figure 1: Parallel corpus annotation and evaluation

Corpus	Description
movie	Monolingual pre-annotated opinion corpus (10K sentences)
movie-p1	Part 1 of *movie* (90%): used to build the classifier which is used for annotation task
movie-p2	Part 2 of *movie* (10%): This is the (test corpus) which is used to test the annotated corpora
parallel	Parallel corpora (183K sentences)
opinion-parallel	Automatically annotated

Table 4: Corpora description

First, a Naive Bayes classifier is built using the movie corpus, then the classifier is used to automatically annotate English sentences of parallel corpus. Then, the same opinion label is assigned to corresponding Arabic sentences of parallel corpus as illustrated in Figure 1. Corpora which are denoted in Figure 1 are described in Table 4.

The Naive Bayes classifier is trained (step 2 in Figure 1) on 3-grams features of each sentence of *movie-p1* corpus. We used 90% of *movie-p1* for training and kept 10% for testing. A sample of subjective/objective n-gram features

[1] Agence France Presse www.afp.com
[2] Annahar News paper www.annahar.com
[3] Assabah newspaper www.assabah.com.tn
[4] National Institute of Technology www.nist.gov
[5] www.ted.com
[6] www.tatoeba.org

subjective	objective
...entertaining...	...the story of...
...interesting...	...order...
...entertainment...	...decides...
...you are...	...discover...
...but it...	...she is...
...me...	...with her...
...fans...	...led...
...it is not...	...kill...
...is so...	...his family...
...I...	...to kill...
...if it...	...one day...

Table 5: A sample of subjective/objective n-gram features of the classifier

of the classifier are presented in Table 5. As can be noted from the table, subjective n-gram features are expressions that the reviewer uses to express what he/she thinks about the movie, while objective n-gram features are used by the reviewer to describe the events of the movie. The accuracy of this classifier on the 10% of *movie-p1* is 0.926 and objective F-Measure ($F1$) is 0.926 and subjective $F1$ is 0.927.
So Naive Bayes classifier is used for automatic annotation task as described in our method steps in Figure 1. In sake of robustness, sentence annotation (steps 3 and 4 in Figure 1) is made only if the probability of the highest class label is above 0.8.

2.3. Annotation Results

The annotation method is applied on the parallel corpus, and results are presented in Table 6. The table presents percentage of annotated sentences with respect to the corpus, the table also shows the class distribution of annotated sentences for each corpus (steps 3 and 4 of Figure 1). As can be seen in the table, 81% of all sentences are annotated, 45% of these are subjective and 55% are objective.

Corpus	Annotated sentences	Subjective	Objective
AFP	90.6%	9.4%	90.6%
ANN	89.9%	18.6%	81.4%
ASB	91.7%	17.8%	82.2%
TED	88.7%	74.8%	25.2%
UN	89.6%	15.7%	84.3%
Medar	89.6%	25.8%	74.2%
NIST	79.4%	20.6%	79.4%
Tatoeba	86.4%	59.3%	40.7%
Total	81%	45%	55%

Table 6: Parallel corpus annotation

A preliminary evaluation (step 6 of Figure 1) of the annotation process consists in: (1) training Naive Bayes classifiers using *parallel-opinion* corpus. (2) testing the obtained classifiers on *movie-p2*. Results are presented in Table 7. The classification accuracy is 0.79, subjective $F1$ is 0.81, and objective $F1$ is 0.74.
To make sure that the classifier built from movie domain can correctly identify the subjectivity of sentences of parallel corpus which is composed of news and non-news domains, a secondary evaluation is done on a sample of 330 sentences selected randomly and annotated by the first author of this paper (step 7 in Figure 1). News sentences are selected from AFP, ANN, ASB, Medar, and NIST corpora, while non-news sentences are selected from TED, UN, Tatoeba corpora. This manually annotated material is described in Table 8. Then, the classifier built from movie reviews is tested on these sentences (step 8 in Figure 1). The results of the classifier on these sentences are described in Table 9. The classification accuracy of news sentences is 0.718, subjective $F1$ is 0.717, and objective $F1$ is 0.720, while the classification accuracy of non-news sentences is 0.658, subjective $F1$ is 0.667, and objective $F1$ is 0.649. As can be seen from the results, the classifier built from movie reviews can classify sentences from other domains: the classifier can detect subjective and objective sentences from news corpus and from non-news corpus. In other words, the classifier does not modelize the genres of the corpus (news/non-news) but the objectivity/subjectivity. Another validation by language models is presented in the next section.

Accuracy	subjective $F1$	objective $F1$
0.79	0.81	0.74

Table 7: Evaluation on *movie-p2* (step 6)

Corpus	Subjective	Objective
News	112	101
Non-News	60	57
Total	330	

Table 8: The sample of parallel sentences annotated by a human (step 7)

Corpus	Accuracy	subjective $F1$	objective $F1$
News	0.718	0.717	0.720
Non-News	0.658	0.667	0.649

Table 9: Evaluation on parallel corpus annotated by a human (step 8)

3. Statistical Language Modelling

In this section, three experiments are conducted in order to show the distinction of subjective/objective text in terms of statistical language modelling. The first one inspects the perplexity of opinionated language models on subjective and objective texts. In the second experiment, we investigate if opinionated language models built from parallel corpora fit the movie domain. In this experience, we compare opinion language models built from movie review domain with language models built from *parallel-opinion* which is composed of UN resolutions, newspapers, and talks. The third experiment inspects the perplexity of opinionated language models on comparable corpora to explore their subjectivity.

3.1. Opinionated Language Models

Regarding to the first experiment, we show that dividing our corpus into two parts (subjective and objective) can be useful for language modelling in terms of perplexity. For that, we measure the perplexity of language models trained on subjective corpus and objective corpus of "*parallel-opinion*".

	Words	Vocabulary
English	4.4M	138K
Arabic	3.7M	264K

Table 10: *parallel-opinion* information

The "*parallel-opinion*" corpus (the output of steps 3 and 4 of Figure 1) is divided into subjective and objective parts according to the automatic labelling. Each one of them is split into 90% for training and 10% for testing. Table 10 describes the number of words and the vocabulary size of the training parts of the corpus. Therefore, we have a subjective training corpus (STrC), an objective training corpus (OTrC), a subjective test corpus (STeC) and an objective test corpus (OTeC). A 3-gram language model is built on STrC (called SLM1) and another one on OTrC (called OLM1). SRILM toolkit (Stolcke and others, 2002) is used to build language models, with Kneser-Ney discounting method. The vocabulary for SLM1 and OLM1 is made up of the union of words of STrC and OTrC, and composed at most of of 138K English words and 264K Arabic words as presented in Table 10. This work has been done for English and Arabic languages.

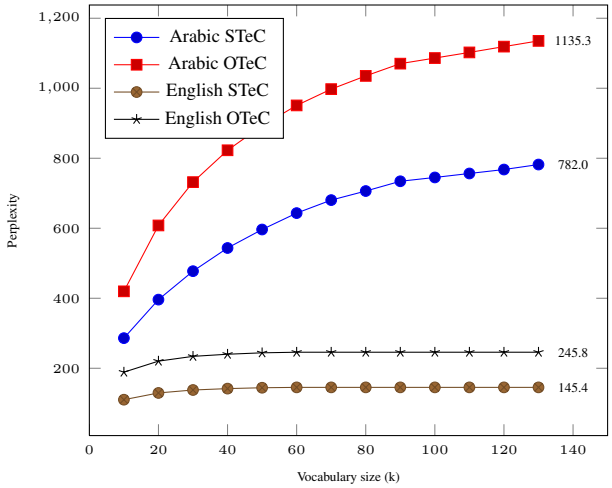

Figure 2: Performance in terms of perplexity of SLM1 on subjective/objective test texts

Figures 2 and 3 present the perplexity of subjective/objective language models (SLM1/OLM1) on subjective STeC and objective OTeC test corpora for several vocabulary sizes. Each vocabulary is made up of the most frequent words. For each figure, the language model is tested on Arabic/English subjective/objective test texts, so we have four performance curves. The numbers on the right side of the figures present the value of the last point of each curve.

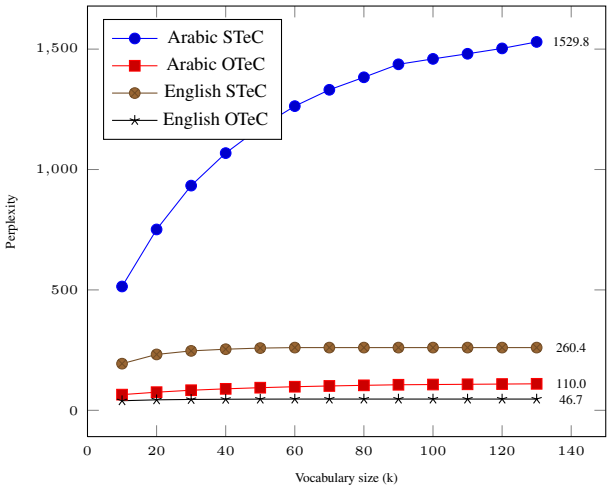

Figure 3: Performance in terms of perplexity of OLM1 on subjective/objective test texts

It can be noted that the perplexity for Arabic is larger than English. According to a study about statistical language modelling for many languages conducted by (Meftouh et al., 2010), Arabic is one of the languages that has high perplexity, because Arabic is agglutinative language and has rich morphology. Second, it can be noted that SLM1/OLM1 models fit better to subjective/objective texts respectively. Indeed, subjective (respectively objective) language models models applied to objective (respectively subjective) test lead to bad performance. It can be concluded from the last result that subjective and objective corpora are very different.

3.2. Testing Language Models

In the second experiment, we inspect the opinion language models built from different topic domain. In this experiment, we test movie review corpus with an opinion language models built from opinion parallel corpora (*parallel-opinion*) which come from different domains (UN resolutions, newspapers, talks). Symmetrically, we test the parallel opinion corpus with opinion language models built from movie corpus. We also aim in this experiment to validate our annotation by testing language models on manually annotated corpus (movie corpus). If subjective/objective models fit better to subjective/objective texts respectively, then our annotation of *parallel-opinion* is good and reliable. We also aim to inspect if the difference between objective and subjective texts is repeated across various topics (movie reviews vs. news).

In this experiment, we have four language models: SLM1/OLM1 of previous section which are trained on subjective/objective parts of *parallel-opinion*, and SLM2/OLM2 which are trained on subjective/objective parts of movie corpus. All models are built using 10K most frequent words vocabulary.

Table 11 gives the perplexity of testing SLM1/OLM1 on movie and *parallel-opinion* test corpus, while Table 12 gives the perplexity of testing SLM2/OLM2 on movie and *parallel-opinion* test corpus. First, we note from the results presented in Tables 11 and 12 that the perplexities

		Models	
Corpus	Test	SLM1	OLM1
Movie	Subjective	368.8	609.7
	Objective	507.8	442.2
parallel-opinion	Subjective	110.2	193.7
	Objective	188.9	39.7

Table 11: Testing models built from *parallel-opinion* corpus

		Models	
Corpus	Test	SLM2	OLM2
Movie	Subjective	354.3	688.9
	Objective	805.2	379.5
parallel-opinion	Subjective	456.3	643.5
	Objective	900.1	687.9

Table 12: Testing models built from *movie* corpus

are higher when models and test corpora are from different domain topics (movie reviews vs. UN resolutions, newspapers, talks). It can be noted also that the perplexity of OLM1 for objective test part of *parallel-opinion* is low. Actually, this comes from the UN corpus, mostly objective, in which numerous sentences contain common parts (for example "taking note of the outcome of the..."); by chance, these common parts are distributed between training and test. We also note that SLM1/OLM1 fit better to subjective/objective parts of movie corpus respectively, and SLM2/OLM2 fit better to subjective/objective parts of *parallel-opinion* respectively. Second, it can be noted that subjective test corpus (respectively objective) does not fit to objective language model (respectively subjective). It can be concluded that language models built from opinion corpora can fit for movie review, namely, the distinction of subjective/objective text is stable across different topic. Additionally, it can be concluded that the annotation of *parallel-opinion* is good and reliable, because subjective (respectively objective) models built from *parallel-opinion* have better perplexity on subjective (respectively objective) corpus manually annotated (the movie corpus). Finally, this experiment confirms the previous results in Section 3.1. which concludes that subjective and objective text are distinct in terms of writing style.

3.3. Testing Comparable Corpora

Regarding the third experiment, opinionated language models built using *parallel-opinion* are inspected on comparable corpora collected by (Saad et al., 2013b). We use two comparable corpus: AFEWC and eNews. AFEWC is collected from Wikipedia, and eNews is collected from euro-news website `http://www.euronews.com`. These corpora are composed of English/Arabic comparable articles aligned at article level. We aim in this experiment to explore the subjectivity of AFEWC and eNews using two ways: annotating with the Naive Bayes classifier, and testing with language models. We also want to show whether the results of annotation and language model tests accord to each others or not. We take a random subset of sentences composed of these corpora for our experiments. The subset size is 30K words.

	Subjective	Objective
AFEWC English	24%	76%
AFEWC Arabic	18%	82%
eNews English	23%	77%
eNews Arabic	11%	89%

Table 13: Subjective/Objective sentences distribution of the subset of the comparable corpora (30K words)

We annotate automatically each sentence in AFEWC and eNews with subjective/objective labels using the Naive Bayes classifier built using *parallel-opinion*. Subjective/objective distribution of eNews and AFEWC comparable corpora is presented in Table 13. The table presents the percentage of subjective and objective sentences in the subset of the comparable corpora. As can be seen from the table, for both comparable corpora, and for both languages, there are more objective sentences than subjective ones. This is coherent because Wikipedia and news mostly tend to be objective: they describe facts.

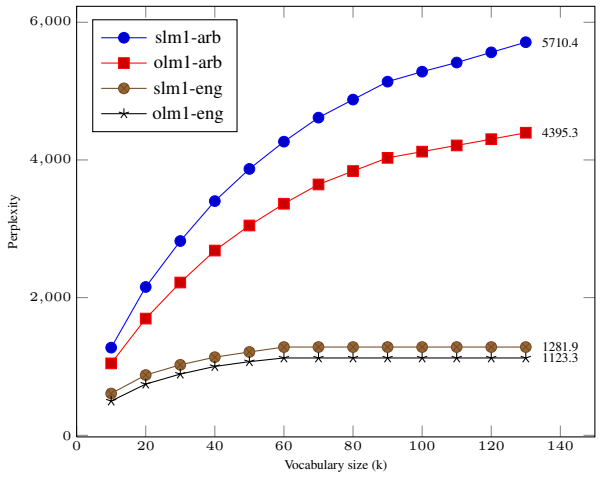

Figure 4: SLM1/OLM1 test on AFEWC

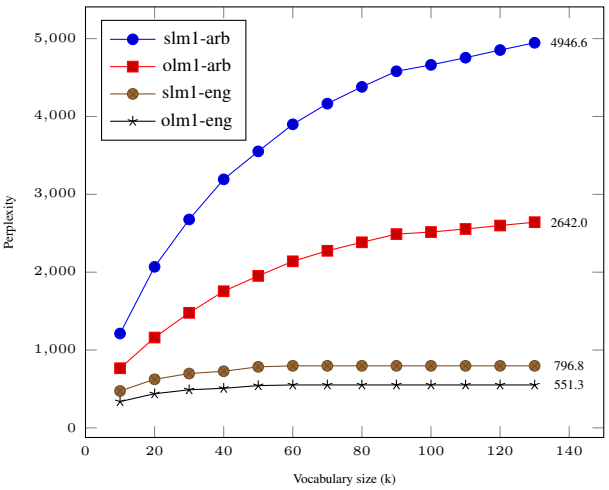

Figure 5: SLM1/OLM1 test on eNews

To confirm this objectivity of the comparable corpora, we

test them with opinion language models SLM1 and OLM1 built from opinion corpora (*parallel-opinion*). Results of these tests are presented in Figures 4 and 5. The Figures present the perplexity of SLM1/OLM1 tested on AFEWC and eNews comparable corpora. We first note that the perplexity values for all test are high. This is maybe because language models are built from different corpus and domain than the test one. We also note from the figures that OLM1 fits better to English/Arabic AFEWC and eNews comparable corpora (has lower perplexity) than SLM1. Classification results and language models tests in the Table 13 and the Figures 4 and 5 confirm the distinction of subjective/objective text. Their results accord to each others. This leads us to confirm that comparable corpora eNews and AFEWC are more objective.

4. Conclusion

We have presented a method for cross-lingual annotation with subjective/objective labels. We tested successfully our classifier with manually labelled data. By training and testing statistical language models, we also showed that subjective/objective texts are statistically different in terms of writing style. We tested several models and corpora of different genres and we always obtained the same conclusion: subjective/objective texts are distinct. Moreover, results of the classifier and of objective/subjective language models highlighted that our comparable corpora (AFEWC and eNews: data extracted from Wikipedia and news website in English and Arabic) are more objective than subjective.
In the future, this work can be extended to polarity in order to go towards a review system which is able to retrieve multilingual articles about a topic, and to classify these articles in terms of opinion.

5. References

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May.

Xiaoyi Ma and Dalal Zakhary. 2009. Arabic newswire english translation collection. Linguistic Data Consortium, Philadelphia.

Karima Meftouh, Med Tayeb Laskri, and Kamel Smaïli. 2010. Modeling Arabic Language using statistical methods. *Arabian Journal for Science and Engineering*, 35(2C):69–82.

Multimodal Information Group NIST. 2010. NIST 2008/2009 open machine translation (OpenMT) evaluation. Linguistic Data Consortium, Philadelphia.

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.

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

Alexandre Rafalovitch and Robert Dale. 2009. United nations general assembly resolutions: A six-language parallel corpus. In *Proceedings of the MT Summit XII*, volume 13, pages 292–299.

Motaz Saad, David Langlois, and Kamel Smaïli. 2013a. Comparing multilingual comparable articles based on opinions. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 105–111, Sofia, Bulgaria, August. Association for Computational Linguistics.

Motaz Saad, David Langlois, and Kamel Smaïli. 2013b. Extracting comparable articles from wikipedia and measuring their comparabilities. *Procedia - Social and Behavioral Sciences*, 95(0):40 – 47. Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013).

Andreas Stolcke et al. 2002. SRILM-an extensible language modeling toolkit. In *Proceedings of the international conference on spoken language processing*, volume 2, pages 901–904.

Jrg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).