UNIVERSITY OF WEST BOHEMIA

FACULTY OF APPLIED SCIENCES

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Multilingual Summarisation and Sentiment Analysis

Habilitation Thesis

Josef Steinberger

April 2013

Summarisation and sentiment analysis are the key NLP technologies which allow monitoring evolving content and opinions in huge amounts of textual data available on the web. Summarisation can address the problem of information overload by extracting and presenting the main content and sentiment analysis can identify opinions expressed towards entities or events. Because there can be found so many opinions, it is needed to aggregate them and present to a user only the most important ones. And this is the case in which summarisation and sentiment analysis have to work together. Studying the problems in multiple languages, besides providing multilingual information access, opens new possibilities, like analysing disagreements in reporting across languages or producing more coherent summaries in the case of weakly covered languages. My research focussed mainly on news data, however, the attention is now shifting towards rising social media. This thesis describes the crossing paths of my research of summarisation and sentiment analysis in multilingual environment.

Content

1 Introduction

Professional and non-professional consumers of news struggle to identify and absorb the relevant information from the overwhelming quantity of news in a multiplicity of different languages currently available on the Web. This is even augmented by the increasing amount of information in social media where mass opinions can be monitored. My research addresses the problem by extracting and presenting a gist of multilingual news and social media content.

Media professionals, however, will want to go beyond summaries from sources in one language and consider how news events are reported in other countries and from other perspectives. At present it does not make sense to attempt fully cross-lingual summarisation, because the multilingual task is challenging enough. However, identifying differences in opinion towards entities and events may provide some clues as the disagreements in reporting across languages, and may also help producing more coherent summaries within each language. I investigated multilingual sentiment analysis which allows, in cooperation with a summariser, generating summaries that reveal these disagreements.

The research aims to push forward the boundaries of summarisation and sentiment analysis research by developing methods that work in a very high volume and a highly multilingual setting.

I worked on three related topics: coreference resolution, summarisation and sentiment analysis[1]. In summarisation, information about corefering expressions can improve content selection and it can be used for correcting entity mentions in a summary, as sentences are extracted without the previous context. In sentiment analysis, coreference resolution is needed to identify a sentiment target. Summarisation and sentiment analysis are joined in the task of opinion summarisation.

The thesis is organised as follows: chapter 2 describes advances in my principal research field which is summarisation. I developed a multilingual summariser which was very successful in evaluation campaigns such as the TAC[2] and thus it can be called state-of-the-art. The next chapter (3) follows the work which has been done for evaluating summaries in multiple languages. Chapter 4 describes a *triangulation* method for creating sentiment dictionaries and evaluation of an entity-centred sentiment analyser on a parallel corpus. Chapter 5 unites the summarisation and sentiment analyses tasks, as it deals with opinion summarisation in social media. The last chapter gives conclusions and current directions of my research.

[1] The research topics are closely related to my postdoc stay at the Europe Media Monitor (EMM) Labs at the Joint Research Centre (JRC) of the European Commission. I worked on new functionalities for EMM which is a web service providing an aggregation and structuring of multilingual online news articles.

[2] The National Institute of Standards and Technology (NIST) initiated the Document Understanding Conference (DUC) series (http://duc.nist.gov/) to evaluate automatic text summarization. Its goal is to further progress in summarization and enable researchers to participate in large-scale experiments. Since 2008 DUC has moved to TAC (Text Analysis Conference) (http://www.nist.gov/tac) that follows the summarization evaluation roadmap with new or upgraded tracks.

2 Language-independent summarisation

SELECTED PAPERS

A. Steinberger, J., Poesio, M., Kabadjov, M. and Ježek, K.: Two Uses of Anaphora Resolution in Summarization. In: Information Processing & Management 43(6), pages 1663-1680, Elsevier. 2007.

B. Steinberger, J., Belyaeva, J., Crawley, J., Della Rocca, L., Ebrahim, M., Ehrmann, M., Kabadjov, M., Steinberger, R. and van der Goot, E.: Highly Multilingual Coreference Resolution Exploiting a Mature Entity Repository. In: Proceedings of the 8th International Conference Recent Advances in Natural Language Processing (RANLP), pages 254-260, Hissar, Bulgaria, 2011.

C. Kabadjov, M., Steinberger, J. and Steinberger, R.: Multilingual Statistical News Summarization. In: Thierry Poibeau, Horacio Saggion, Jakub Piskorski & Roman Yangarber (eds.), Multi-source, Multilingual Information Extraction and Summarization, pages 229-252, Springer, 2013.

Automatic summarisation deals with the problem of producing a succinct informative gist for a document (or a set of documents about the same topic). The aim of the task could be that the target language of the summary be the same as the input documents (standard single-/multi-document summarization) (Nenkova and Louis, 2008) or that the languages of summary/input documents be different (cross-language document summarization) (Wan et al., 2010). Moreover, the task of handling several languages, with summary and input documents being in the same language, has been termed as multilingual summarization (Litvak et al., 2010). If the summariser does not use any language-specific resources or properties we can it language-independent summarisation.

Summarization has been an active area of research for several decades (Luhn, 1958; Edmundson, 1969), but in particular over the past seventeen years. The area was initially focused on singledocument summarization (Mani and Maybury, 1999), a fact reflected by the first US NIST's Document Understanding Conference (DUC) evaluation exercises (Over et al. 2007). Then, over the past decade the emphasis shifted to multi-document summarization exemplified by latter DUCs followed by the Text Analysis Conference (TAC) exercises. However, it has been only until recently that interest in multilingual summarization has risen (Kabadjov et al., 2009; Litvak et al., 2010).

In this chapter, I follow my research on the way from single-document summarisation, via multi-document summarisation, to the final goal: multilingual multi-document summarisation. In section 2.1, I describe the basic single-document summarisation approach and the way how anaphora resolution (identifying successive entity mentions) can improve content selection and summary coherence. The proposed approach based on Latent Semantic Analysis (LSA) can work with more input documents. However, intra-document coreference (solved by anaphora resolution in the single-document scenario) has to be extended to inter-document coreference (discussed in section 2.2). Thanks to language-independence of LSA and multilingual properties of the coreference resolver, the approach can be applied to multiple languages (section 2.3). At the end of this chapter I summarise our participation at the TAC evaluations (2008-2011).

2.1 Single-document summarisation and anaphora resolution

Information about anaphoric relations could be beneficial for applications such as summarization and segmentation, which involve extracting discourse models (possibly very simplified) from text. In this work (Steinberger et al., 2007) we investigated exploiting automatically extracted information about the anaphoric relations in a text for two different aspects of the summarization task. First of all, we used anaphoric information to enrich the latent semantic representation of a document (Landauer and Dumais, 1997), from which a summary is then extracted. Secondly, we used anaphoric information to check that the anaphoric expressions contained in the summary thus extracted still have the same interpretation that they had in the original text.

2.1.1 LSA-based summarization

LSA (Latent Semantic Analysis (Landauer and Dumais, 1997)) is a technique for extracting the 'hidden' dimensions of the semantic representation of terms, sentences, or documents, on the basis of their use. It has been extensively used in educational applications such as essay ranking (Landauer and Dumais, 1997), as well as in NLP applications including information retrieval (Berry et al., 1995) and text segmentation (Choi et al., 2001). In 2002, a method for using LSA for summarization has been proposed in (Gong and Liu, 2002). This purely lexical approach was the starting point for our own work. We changed the selection criterion to include in the summary the sentences which have greatest combined weight across all important topics (dimensionality is reduced to r dimensions). After placing a sentence in the summary, the topic/sentence distribution is changed by subtracting the information contained in the selected sentence. The vector lengths of similar sentences are decreased, thus preventing within summary redundancy. For details see (Steinberger and Ježek, 2009).

2.1.2 Using anaphora resolution for content selection

Identifying central characters is crucial to provide a good summary. Among the clues that help us to identify such 'main characters,' the fact that an entity is repeatedly mentioned is clearly important. Methods that only rely on lexical information to identify the main topics of a text can only capture part of the information about which entities are frequently repeated in the text. What anaphora resolution can do for us is to identify which discourse entities are repeatedly mentioned, especially when different forms of mention are used. We can then use the anaphoric chains identified by the anaphoric resolvers as additional terms in the initial LSA matrix.

The anaphora resolution system we used, GUITAR (Kabadjov, 2007), is a publically available tool designed to be modular and usable as an off-the-shelf component of a NLP pipeline. It can resolve pronouns, definite descriptions and proper nouns. We evaluated both the lexical and the anaphoric+lexical summarizers using the DUC2002 corpus and the ROUGE measure (Lin, 2004), which made it easier to contrast our results with those published in the literature. We found that the incorporated anaphoric knowledge improved the summariser which performed on the level of the state-of-the-art systems.

2.1.3 Using anaphora resolution for checking entity references in a summary

Anaphoric expressions can only be understood with respect to a context. This means that summarization by sentence extraction can wreak havoc with their interpretation: there is no guarantee that they will have an interpretation in the context obtained by extracting sentences to form a summary, or that this interpretation will be the same as in the original text. Another use for anaphora resolution in summarization is correcting the references in the summary. Our idea was to replace anaphoric expressions with a full noun phrase in the cases where otherwise the anaphoric expression could be misinterpreted. In the task of correcting entity mentions in a summary we observed precision 69%.

Details can be found in (Steinberger et. al, 2007) – Appendix A.

2.2 Inter-document cross-lingual coreference resolution

Recent work on coreference resolution has been largely dominated by machine learning approaches and predominantly for the English language in great part due to the availability of annotated corpora. We addressed two important remaining gaps in coreference resolution. Firstly, we were interested in highly multilingual coreference. Secondly, we addressed the problem of common noun coreference by exploiting a large lexical resource, the named entity database, compiled over the past few years by automatically extracting names from hundreds of thousands of online news articles in twenty languages. The coreference resolver we presented was designed to work as part of the Europe Media Monitor (EMM) system[3].

2.2.1 The multilingual named entity database

The historical repository of EMM's person and organization titles is a by-product of the Named Entity Recognition (NER) process, which has been applied daily to tens of thousands of multilingual news articles per day since 2004 (Pouliquen and Steinberger, 2009). Titles are parts of the

[3] http://emm.newsbrief.eu/overview.html

name recognition patterns, and each time a name is found, EMM keeps track of the titles found next to the name. The result is a large multilingual repository of titles and other attributes about names.

2.2.2 The coreference algorithm

The coreference resolution module is built for inclusion in a larger pipeline architecture, where an input text document undergoes several processing phases during which the source is augmented with layers of meta data such as named entities. Before running the coreference resolution module, known and guessed entities are found in the text. Known entities are entities that have been found in at least five different news clusters in the past. The entity guessing part identifies previously unseen entities. The coreference resolver links mentions (name parts or title/function references) to entities using the reference-entity associations obtained by querying the named entity database.

In order to evaluate our coreference system we compiled a corpus of news articles in seven different languages: English, German, Italian, Spanish, French, Russian and Arabic, thus, covering a diverse set of language family branches as are Germanic, Romance, Slavic and Semitic. Not surprisingly, the overall coreference resolution of proper names yields high precision (98%). What is more significant, however, is the performance on person titles, which entail mostly references by means of definite descriptions not sharing a head noun with the antecedent, where the system surpasses the 70% threshold (with the exception of French with 61.2%). It is worth pointing out that these are largely regarded as among the most challenging to resolve, mainly because their resolution requires real-world knowledge.

Details can be found in (Steinberger et. al, 2011c) – Appendix B.

2.3 Multilingual multi-document summarisation

The building stones for our multilingual multi-document summariser are language-independent LSA (described in section 2.1.1) and the coreference resolution (described in 2.2). The LSA approach is similar to the single-document approach, however, the input term-by-sentence matrix is built from all sentences in a set of documents. The sentence selection step protects from extracting redundant content. As in the case of single-document summarisation and anaphora resolution, the notion of term is generalised. In addition to words, it uses mentions of discourse entities, thus enhancing the original lexical LSA summarization. In this case, however, mentions of the same entity have to be linked across the border of a document which leads to using inter-document coreference resolution, e.g. as it was described in section 2.2. Augmenting the initial matrix with information about disambiguated entities naturally provides not only stronger inter-sentential cohesion (i.e., the LSA clusters sentences from different documents that make reference to the same entities), but also provides multilingual capabilities inherited by the multilingual entity disambiguation. Thus, this approach to summarization is not only multi-document, but also multilingual.

Details can be found in (Kabadjov et al., 2013) – Appendix C.

2.3.1 Participation in TACs

We participated in all editions of Text analysis conference (TAC) evaluation campaigns.

We started in 2008 with the lexical LSA-based approach, which tries to capture, and extract the best sentences about, the most important concepts in the source articles, as described in section 2.1.1. The system was ranked 9th in overall responsiveness within the 58 participating systems (Steinberger and Ježek, 2009a).

In 2009, we included named entities in the summarizer's input representation, as described in section 2.2.2 and it resulted in 2nd place among 52 runs (Steinberger et al., 2010a).

TAC 2010 encouraged an even deeper semantic analysis of the source documents by its new Guided summarization task. The systems were given a list of aspects for each article category (e.g. for category *attacks*: *what happened, where, when, why, perpetrator, who affected, damages, countermeasures*), and the summary should include those aspects if possible. Per-category aspects that should guide the summarizer were identified by an event-extraction system and automatically generated lists of terms semantically related to the predefined aspects. It extracted sentences which contained the most important concepts of LSA and also relevant aspects (Steinberger et al., 2011a).

We participated with the summarizer, improved by temporal analysis and sentence (re-)generation approach (Turchi et al., 2010), also in the next campaign (TAC'11). In the Guided summarization task (English news clusters), our system was ranked high in linguistic quality and above average in content. We were very successful in the new multilingual summarization task, in which our system performed best in 5 from 7 languages (Steinberger et al., 2012a).

3 Summarisation evaluation in multiple languages

SELECTED PAPERS

D. Giannakopoulos, G., El-Haj, M., Favre, B., Litvak, M., Steinberger, J. and Varma, V.: TAC 2011 multiling pilot overview. In: Proceedings of the Text Analysis Conference 2011, National Institute of Standards and Technology (NIST), Gaithersburg, USA, 2011.

E. Turchi, M., Steinberger, J., Kabadjov, M. and Steinberger, R.: Using Parallel Corpora for Multilingual (Multi-Document) Summarisation Evaluation. In: Multilingual and Multimodal Information Access Evaluation, Springer Lecture Notes for Computer Science 6360, pages 52-63, Springer, 2010.

F. Steinberger, J. and Turchi, M.: Machine Translation for Multilingual Summary Content Evaluation. In: Proceedings of the NAACL Workshop on Evaluation Metrics and System Comparison for Automatic Summarization, pages 19-27, Montreal, Canada, ACL, 2012.

Evaluation of automatically produced summaries in different languages is a challenging problem for the summarization community, because human efforts are multiplied to create model summaries for each language. Unavailability of parallel corpora suitable for news summarization adds even another annotation load because documents need to be translated to other languages. At the TAC'11 campaign, six research groups spent a lot of work on creating evaluation resources in seven languages (Giannakopoulos et al., 2012). Thus compared to the monolingual evaluation, which requires writing model summaries and evaluating outputs of each system by hand, in the multilingual setting we need to obtain translations of all documents into the target languages, write model summaries and evaluate the peer summaries for all the languages. I describe the TAC'11 multilingual task (section 3.1), in which I took the lead of the organisation of the Czech language subtask (translation, annotation and evaluation of the participated systems). Then I describe ideas we proposed to do the evaluation automatically: creating a parallel corpus and projecting the annotation (section 3.2) and using machine translation (section 3.3).

3.1 Community evaluation effort – TAC'11 Multiling

The *Multiling* Pilot introduced in TAC 2011 was a combined community effort to present and promote multi-document summarization approaches that are (fully or partly) language-neutral. To support this effort an organizing committee across more than six countries was assigned to create a multilingual corpus on news texts, covering seven different languages: Arabic, Czech, English, French, Greek, Hebrew and Hindi. Our responsibility was creation of the Czech subcorpus and then evaluation of summaries submitted by systems participating in this shared task.

Overall, the task was successful, although the costs were enormous. A corpus of 10 topics (each topic contained 10 English articles) was created, and translated to the rest of the languages, and 3 summaries per topic were manually written for all the 7 languages. It bootstrapped multilingual summarization research as a community effort, by bringing together researchers from a variety of institutions and countries, aiming to tackle the same problem. It provided a method and an estimated cost for the creation of a multilingual summarization dataset. It created such a benchmark dataset in 7 languages, using openly and freely available texts. The dataset is itself provided freely, upon request. It indicated that there exist systems that perform good-enough summarization in several languages (our LSA-based summariser performed the best in 5 from 7 languages).

For details see Appendix D (Giannakopoulos et al., 2012). The shared task with continue this year as a workshop of ACL'2013.

3.2 Using parallel corpora

Because of the huge cost of creating a multilingual summarisation corpus we tried to develop a method which would not require manual translation (the most expensive part of the TAC Multiling) and writing summaries for all investigated languages.

Parallel corpora – texts and their exact translation – are widely used to train and evaluate statistical machine translation systems. Some of the most widely known freely available parallel corpora are Europarl (Koehn, 2005) and JRC-Acquis (R. Steinberger et al., 2006). Many of them are from the domains of law and public administration and they are not suitable for summarisation evaluation. This was the reason why we used articles from the Project Syndicate website[4]. Given a set of parallel and sentence-aligned documents in several languages referring to a particular topic (a document cluster), our approach consists of manually selecting the most representative sentences in one of the languages (the pivot language). This sentence selection is then projected to all the other languages, by exploiting the parallelism property of the documents. The result is a multilingual set of sentences that can be directly used to evaluate extractive summarisation. When several annotators select sentences, the sentences can additionally be ranked, depending on the number of annotators that have chosen them.

[4] http://www.project-syndicate.org/. Project Syndicate is a voluntary, member-based institution that produces high quality commentaries and analyses of important world events. Each contributor produces a commentary in one language. This is then human-translated into various other languages.

Details of the experiment can be found in Appendix E (Turchi et al., 2010). This work was done before the TAC Multiling. The advantage of our approach via parallel corpus and sentence selection is much lower cost. Disadvantage compared to the TAC Multiling are that it can evaluate only sentence-extractive summarisation as no human-created summaries are available.

For details see Appendix E (Turchi et al., 2010).

3.3 Using machine translation

In both approaches described in the previous sections, manual annotation was needed as TAC approach is completely manual and using the parallel corpus and sentence projection is semi-automatic. We investigated whether machine translation could be useful for the summarisation evaluation task.

In the last fifteen years, research on Machine Translation (MT) has made great strides allowing human beings to understand documents written in various languages. Nowadays, on-line services such as Google Translate and Bing Translator[5] can translate text into more than 50 languages showing that MT is not a pipe-dream. We used our in-house translation service (Turchi et al., 2012) which is based on the most popular class of Statistical Machine Translation systems (SMT): the Phrase-Based model (Koehn et al., 2003).

We analysed whether the three manual parts of TAC Mutliling (translation of articles, writing summaries for each language and evaluating system summaries) can be automatised: automatically translate the English articles and manual summaries to the other languages and evaluate system summaries by ROUGE.

Our results showed that quality of machine translation has not reached the level to test behaviour of summarisers on translated articles. The use of translated models (manual summaries) does not alter much the overall system ranking. It maintained a fair correlation with the source language ranking although without statistical significance in most of the cases given the limited data set. A drop in ROUGE score was evident, and it strongly depended on the translation performance. Using automatic evaluation methods like ROUGE instead of manual evaluation is discussed in the community extensively. ROUGE is widely used because of its simplicity and its high correlation with manually assigned content quality scores on overall system rankings, although per-case correlation is lower.

The study (Steinberger and Turchi, 2012b) left many opened questions: What is the required translation quality which would let us substitute target language models? Are translation errors averaged out when using translated models from more languages? Can we add a new language to the TAC multilingual corpus just by using MT having in mind lower quality (lower scores) and being able to quantify the drop? Experimenting with a larger evaluation set could try to find the answers.

For details see Appendix F (Steinberger and Turchi, 2012b).

[5] http://translate.google.com/ and http://www.microsofttranslator.com/.

4 Entity-centred multilingual sentiment analysis

SELECTED PAPERS

G. Steinberger, J., Ebrahim, M., Ehrmann, M., Hurriyetoglu, A., Kabadjov, M., Lenkova, P., Steinberger, R., Tanev, H., Vázquez, S., Zavarella, V.: Creating sentiment dictionaries via triangulation, In: Decision Support Systems 53, pages 689–694, Elsevier, 2012.

H. Steinberger, J., Lenkova, P., Kabadjov, M., Steinberger, R. and van der Goot, E.: Multilin-gual Entity-Centered Sentiment Analysis Evaluated by Parallel Corpora. In: Proceedings of the 8th International Conference Recent Advances in Natural Language Processing, pages 770-775, Hissar, Bulgaria, 2011.

In sentiment analysis the goal is to detect and classify subjective content of a text. The text can be classified as a whole such as in product reviews, in which an overall judgment is assigned to the product. If we move to the news domain, the overall sentiment score of an article can be used for detecting bad or good news. It can be used also for detecting the changes in sentiment in a particular topic. However, if the goal is to detect sentiment expressed towards entities, the aggregated sentiment of the articles, in which the entity appears, need not to correspond to opinions expressed towards the entity. The entity can be mentioned positively in a very negative article. We have to go down and analyze each entity mention based on the surrounding context. Solving the problem in multilingual environment and gathering large amounts of articles from many sources give advantage to detect news opinions expressed in different countries towards same persons. Also, it eliminates the biased news. However, multilinguality brings another challenge. For instance, it is not easy to develop NLP tools like parsers or taggers in many languages, also using them can cause computational problems when applied on large amounts of articles every day. Another difficulty comes with resources. Sentiment-annotated data are not usually available for other types of texts then reviews, or they are almost exclusively available for English. Sentiment dictionaries are also mostly available for English only or, if they exist for other languages, they are not comparable, in the sense that they have been developed for different purposes, have different sizes, are based on different definitions of what sentiment or opinion means.

We addressed the resource bottleneck for sentiment dictionaries, by developing highly multilingual and comparable sentiment dictionaries having similar sizes and based on a common specification (section 4.1 – Steinberger et al., 2012c). Our sentiment system is simply based on counting subjective terms around entity mentions (mainly persons and organizations). Evaluating its performance in more languages would multiply the annotation efforts. In section 4.2 I describe using parallel corpora to automatically project annotations from English. We studied the subjectivity of the entity-centred sentiment annotation and evaluated our sentiment system in seven languages (English, Spanish, French, German, Czech, Italian and Hungarian). As a side effect this evaluation served as a task-based evaluation of the quality of the sentiment dictionaries.

4.1 Creating sentiment dictionaries via triangulation

The sentiment dictionaries, currently available in 15 languages, were created using a triangulation method, which was described in detail in (Steinberger et al., 2012c). In a nutshell: carefully elaborated English and Spanish sentiment word lists were translated into third languages. The introduction of errors through word sense ambiguity was limited by taking the intersection of both target language word lists. According to our evaluation, approximately 90% of these intersection words were correct, while only about 50% of those words were correct that were translations from either English or Spanish, but not from both. For Arabic, Czech, French, German, Italian and Russian, these word lists were manually checked and enhanced, while for Bulgarian, Dutch, Hungarian, Polish, Portuguese, Slovak and Turkish we simply used the intersecting word list. For a subset of languages (Czech, English and Russian), wild cards were manually added to the sentiment word lists in order to capture morphological variants.

For details see appendix G (Steinberger et al., 2012c).

4.2 Entity-centred SA system

Our objective was to detect positive or negative opinions expressed towards entities in the news across different languages and to follow trends over time. Entities of interest are mostly persons and organisations, but also concepts such as the ’7th Framework Program’ or ’European Constitution’. Entities can be mentioned positively in negative news context, and vice versa, so that document level analysis is not sufficient (Balahur et al., 2010), but opinions expressed towards the specific entity mention must be detected. As we do not have access to parsers or even part-of-speech taggers for the range of languages we intend to analyse, we chose to use an extremely simple method that does not require language-specific tools besides NER software and language-specific sentiment dictionaries: we add up positive and negative sentiment scores in six-word windows around the entities, distinguishing two positive and two negative levels of sentiment words (having values of -4, -2, 2 and 4 points, respectively). Enhancers and diminishers add or remove 1 point, negation inverts the value, except for negated high positive (‘not very good’ is not equivalent to ‘very bad’).

For details see Appendix H (Steinberger et al., 2011b).

4.3 Evaluation by parallel corpora

We worked with data from Workshops on Statistical Machine Translation (2008, 2009, 2010)[6] which provide parallel corpora of news stories in 7 European languages: English, Spanish, French, German, Czech, Italian (only 2009) and Hungarian (only 2008 and 2009). Putting together the data from the three years resulted in 7 065 parallel sentences in five languages, and a subset in Italian and Hungarian. We ran our in-house entity recognition on the data. Only known entities (entities present in our database) were marked in the data. It gave us enough samples to run sentiment experiments although guessing other entities (and considering coreference mentions) would considerably increase the pool of samples (see section 2.2). For English we received 1 274 entity mentions, resulting in the same number of sentence-target (S-T) pairs for testing sentiment analysis. Because of different performance of entity recognition we obtained fewer S-T pairs in other languages than in English.

We built golden standard annotations in English. Then we projected the sentiment polarities in golden standard data to other languages and we ran the sentiment system. The results showed the overall agreement with golden standard from 66% (Italian) to 74% (English and Czech). The best two performing languages were the ones with all steps of dictionary creation finished, showing that the evaluation also serves as a task-based evaluation for sentiment dictionaries. Another observation was that the system performed better on negative statements than on positive ones indicating that negative statements are more explicitly expressed in news. Even if discovering the right polarity of sentiment towards an entity in a sentence was a difficult task and the system's results for non-neutral cases were modest, per-entity sentiment aggregation led to precise conclusions when used carefully.

For details see Appendix H (Steinberger et al., 2011b).

[6] http://www.statmt.org/wmt10/translation-task.html.

5

Opinion summarisation

SELECTED PAPERS

I. Balahur, A., Kabadjov, M., Steinberger, J., Steinberger, R. and Montoyo, A.: Challenges and solutions in the opinion summarization of user-generated content. In: Journal of Intelligent Information Systems 39(2), pages 375-398, Springer, 2012.

Recent years have marked the beginning and expansion of the Social Web, which is characterized by the high quantity of user-generated content. The data produced by users has proven useful in many domains (e.g. marketing studies, business intelligence). The high quantity of user-generated content and its proven importance has led to the development of new tasks within NLP that deal with extracting knowledge from the information produced by users. One of these tasks is sentiment analysis, which is also called by some authors *opinion mining*. Sentiments can be present in text directly, indirectly, e.g. by mentioning the a positive or negative effect and implicitly, i.e. through expressions of appraisal, presentation of a affective states, or the indirect mentions of situations which the reader can interpret and to which they can assign an emotional label. Much research in the past years has concentrated on developing systems that deal with sentiment analysis, from a multitude of perspectives, in different languages and from distinct textual genres (e.g. blogs, newspaper articles, forums, reviews). Nevertheless, real-world applications of sentiment analysis often require more than an opinion mining component. In many cases, even the result of the opinion processing by an automatic system still contains large quantities of information, which remain difficult to deal with manually. For example, for questions such as "*Why do people like George Clooney?*" we can find thousands of answers on the Web. Therefore, finding the relevant opinions expressed on George Clooney, classifying them and filtering only the positive opinions is not helpful enough for the users. They will still have to sift through thousands of text snippets, containing relevant, but also much redundant information. Moreover, when following the comments on a topic posted on a blog, for example, finding the arguments given in favour and against the given topic might not be sufficient to a real user. They might find the information truly useful only if it is structured and has no redundant pieces of information. Therefore, apart from analyzing the opinion in text, a real-world application for sentiment analysis should also contain a summarization component.

In (Balahur et al., 2012) we studied the manner in which opinions can be summarized, so that the obtained summary can be used in real-life applications e.g. marketing, decision-making. We discussed the aspects involved in this task and the challenges it implies, in comparison to traditional text summarization, demonstrating how and why it is different from content-based summarization. We compared and evaluated the results of employing opinion mining versus summarization as a first step in opinion summarization.

The main objective of our experiments was to design a system that is able to produce opinion summaries, for two different types of texts: a) blog threads, in which case we aimed at producing summaries of the positive and negative arguments given on the thread topic; and b) reviews, in the context of which we assessed the best manner to use opinion summarization in order to determine the overall polarity of the sentiment expressed. In our first opinion summarization experiments, we adopted a standard approach by employing in tandem a sentiment classification system and a text summarizer. The output of the former was used to divide the sentences in the blog threads into three groups: sentences containing positive sentiment, sentences containing negative sentiment and neutral or objective sentences. Subsequently, the positive and the negative sentences were passed on to the summarizer separately to produce one summary for the positive posts and another one for the negative ones. We used the LSA-based summariser discussed in section 2. The reason for passing the positive and negative sentences separately is that in this manner we ensure that the final summaries will contain opinions from both positive and negative classes. In the opposite case, the summarization system can either choose only sentences expressing positive sentiment, or sentences expressing negative sentiment. The experiments showed that in the case of opinion summarization, performing the summarization step first can lead to the loss of information that is vital from the opinion point of view (i.e. that contains only factual information, and is not useful for an opinion-based summary). Although much remains to be done, the approaches we proposed obtained encouraging results and pointed to clear directions in which further improvements can be made.

For further detail see Appendix I (Balahur et al., 2012).

6 Conclusions and current research

In this thesis I summarised my research in interconnected fields of coreference resolution, summarisation and sentiment analysis.

In summarisation, an LSA-based framework was proposed. Special attention was given to language-independence and to the role which entities play in summarisation. The research started with intra-document coreference (resolved by anaphora resolution) and single-document summarisation, and reached more complicated inter-document and cross-language coreference and multilingual multi-document summarisation. The summariser represents state-of-the-art in multilingual summarisation. While it works with only one language at time, it may not gather enough statistical information about feature co-occurrence in the case of weakly covered languages. I will try to find novel ways of using information from all languages in order to support content selection in target language within the LSA-based framework.

However, automatic summaries are still far from those produced by humans, mainly because they simply select the most important sentences from the source and omit summary generation (sentence compression/combination/rephrasing). We proposed an unsupervised and language-independent approach to summary generation from summary representation based on the LSA framework and on a machine-translation-inspired technique for sentence reconstruction (Steinberger et al, 2010b). The problem is very challenging and finding more features is necessary.

To make summaries more responsive, I also studied guided summarisation, in which the summariser is guided by topic-specific aspects. This leads to using an information extraction tool for filling the aspect slots, which can be then exploited by the summariser (Steinberger et al., 2011d). I will study different ways of using the extracted slots. In (Steinberger et al., 2011d), important sentences that contained the aspects were extracted, however, it is possible to completely generate new sentences by a predefined template summary, or extract only sentence fragments covering the aspects by compressing/regenerating the sentences in which they were contained.

Summarisation evaluation in different languages is a challenging problem for the summarization community, because human efforts are multiplied for each language. I co-organised the TAC'11 community effort to create multilingual summarisation evaluation resources and to assess the quality of state-of-the-art systems. I proposed two ways how to lower the annotation costs: to use a parallel corpus and automatically project sentence annotations and to use a machine translator. A larger corpus is needed to get statistically significant results. I will continue in the community multilingual corpus creation effort which initiated at TAC'11.

In sentiment analysis, I presented the framework for entity-centred multilingual sentiment analysis. The novel triangulation approach used for creating sentiment dictionaries in different languages was invented. An unsupervised sentiment analyser was proposed and evaluated by a parallel corpus. There is much room for improving the sentiment dictionaries. I will try to find a way how to extend them automatically by using semantic spaces or distributional semantics. Gathering opinions in social media seems to be more and more important. Projecting automatically the sentiment dictionaries to the language used in social media would be valuable. For specific tasks like sentiment analysis in Czech social media I will participate in creating a corpus which could be used then to train machine learning approaches. It will be interesting to study both unsupervised and supervised approaches.

I presented a study of opinion summarisation. This is a very actual problem because thousands of opinions about an entity or an event can be found in social media. Analysing their polarity is not enough. A real-world application for sentiment analysis should also highlight the most important (~frequent) opinions, which naturally leads to including a summarization component.

PAPERS ACCEPTED, SUBMITTED OR IN PREPARATION

Steinberger, J., Turchi, M., Tanev, H. and Steinberger, R.: Versatile Summarisation of Multilingual World News. To appear in: Alessandro Fiori (ed.): Innovative Document Summarization Techniques: Revolutionizing Knowledge Understanding, IGI Global, 2013.

Steinberger, J., Kabadjov, M. and Poesio, M.: Coreference Applications to Summarization. To appear in: Massimo Poesio, Roland Stuckardt and Yannick Versley (eds.), Anaphora Resolution: Algorithms, Resources, and Evaluation, Springer, 2013.

Habernal, I., Ptáček, T. and Steinberger J.: Sentiment Analysis in Czech Social Media Using Supervised Machine Learning. Submitted to HLT-NAACL/WASSA'13, ACL, 2013.

Cantarella, S. and Steinberger, J.: Multilingual Index Terms for Information Access: an Approach via Triangulation. Submitted to the 17th Conference on Theory and Practice of Digital Libraries, Springer, 2013.

Tanev, H. and Steinberger, J.: Event Extraction in Slavonic languages. In preparation for ACL/BSNLP'13, ACL, 2013.

References

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., Pouliquen, B., and Belyaeva, J. (2010): Sentiment analysis in the news. In: Proceedings of LREC'10, ELRA.

Balahur, A., Kabadjov, M., Steinberger, J., Steinberger, R. and Montoyo, A. (2012): Challenges and solutions in the opinion summarization of user-generated content. In: Journal of Intelligent Information Systems 39(2), p. 375-398, Springer.

Choi, F. Y. Y., Wiemer-Hastings, P. and Moore, J. D. (2001): Latent semantic analysis for text segmentation. In: Proceedings of EMNLP, Pittsburgh, US, ACL.

Gong, Y. and Liu, X. (2002): Generic text summarization using relevance measure and latent semantic analysis. In: Proceedings of ACM SIGIR, New Orleans, US, ACM.

Edmundson, H. (1969): New methods in automatic extracting. In: Journal of the Association for Computing Machinery 16(2), p. 264–285, ACM.

Giannakopoulos, G., El-Haj, M., Favre, B., Litvak, M., Steinberger, J. and Varma, V. (2012): TAC2011 Multiling pilot overview. In: Proceedings of the Text Analysis Conference 2011, National Institute of Standards and Technology (NIST). Gaithersburg, USA.

Kabadjov, M. (2007): A comprehensive evaluation of anaphora resolution and discourse-new recognition. Ph.D. thesis, Department of Computer Science, University of Essex.

Kabadjov, M., Steinberger, J., Pouliquen, B., Steinberger, R., Poesio, M. (2009): Multilingual statistical news summarisation: Preliminary experiments with english. In: Proceedings of the Workshop on Intelligent Analysis and Processing of Web News Content at the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WIIAT), ACM.

Kabadjov, M., Steinberger, J. and Steinberger, R. (2013): Multilingual Statistical News Summarization. In: Thierry Poibeau, Horacio Saggion, Jakub Piskorski & Roman Yangarber (eds.), Multi-source, Mul-tilingual Information Extraction and Summarization, pages 229-252, Springer.

Koehn, P., Och, F.J. and Marcu, D. (2003): Statistical phrase-based translation. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, p, 48–54, ACL.

Koehn, P. (2005): Europarl: A Parallel Corpus for Statistical Machine Translation. In: X Machine Translation Summit, p. 79–86, Phuket, Thailand.

Landauer, T. K. and Dumais, S. T. (1997): A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. In: Psychological Review, 104, p. 211–240, APA.

Lin, C.Y. (2004): ROUGE: a package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out, Barcelona, Spain, ACL.

Litvak, M., Last, M., Friedman, M. (2010): A new approach to improving multilingual summarization using a genetic algorithm. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, p. 927–936, ACL.

Luhn, H. (1958): The automatic creation of literature abstracts. In: IBM Journal of Research and Development 2(2), p. 159–165, IBM.

Mani, I., Maybury, M. (1999): Advances in Automatic Text Summarization, MIT Press.

Nenkova, A., Louis, A. (2008): Can you summarize this? identifying correlates of input difficulty for generic multi-document summarization. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, p. 825–833. ACL.

Over, P., Dang, H., Harman, D. (2007): DUC in context. In: Information Processing and Management 43(6), p. 1506–1520, Elsevier.

Pouliquen, B. and Steinberger, R. (2009): Automatic construction of multilingual name dictionaries. In: C. Goutte, N. Cancedda, M. Dymetman, G. Foster (eds.) Learning Machine Translation, MIT Press, NIPS series.

Steinberger, R. and Pouliquen, B. and Widiger, A. and Ignat, C. and Erjavec, T. and Tufis, D. and Varga, D. (2006): The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. LREC, p. 24–26. Genova, Italy.

Steinberger, J., Poesio, M., Kabadjov, M.A., Ježek, K. (2007): Two uses of anaphora resolution in summarization. In: Information Processing and Management 43(6), p. 1663–1680, Elsevier.

Steinberger, J. and Ježek, K. (2009a): SUTLER: Update Summarizer Based on Latent Topics. In: Proceedings of the Text Analysis Conference 2008, National Institute of Standards and Technology. Gaithersburg, USA.

Steinberger, J., Ježek, K. (2009b): Update summarization based on novel topic distribution. In: Proceedings of the 9th ACM DocEng, Munich, Germany, ACM.

Steinberger, J., Kabadjov, M., Pouliquen, B., Steinberger, R. and Poesio, M. (2010a): WB-JRC-UT's Participation in TAC 2009: Update Summarization and AESOP Tasks. In: Proceedings of the Text Analysis Conference 2009, National Institute of Standards and Technology. Gaithersburg, USA.

Steinberger, J., Turchi, M., Kabadjov, M., Cristianini, N. and Steinberger, R. (2010b): Wrapping up a Summary: from Representation to Generation. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 382-386, ACL.

Steinberger, J., Tanev, H., Kabadjov, M. And Steinberger, R. (2011a): JRC's Participation in the Guided Summarization Task at TAC 2010. In: Proceedings of the Text Analysis Conference 2010, National Institute of Standards and Technology (NIST). Gaithersburg, USA.

Steinberger, J., Lenkova, P., Kabadjov, M., Steinberger, R. and van der Goot, E. (2011b): Multilingual Entity-Centered Sentiment Analysis Evaluated by Parallel Corpora. In: Proceedings of the 8th International Conference Recent Advances in Natural Language Processing, p. 770-775. Hissar, Bulgaria.

Steinberger, J., Belyaeva, J., Crawley, J., Della Rocca, L., Ebrahim, M., Ehrmann, M., Kabadjov, M., Steinberger, R. and van der Goot, E. (2011c): Highly Multilingual Coreference Resolution Exploiting a Mature Entity Repository. In: Proceedings of the 8th International Conference Recent Advances in Natural Language Processing, p. 254-260, Hissar, Bulgaria.

Steinberger, J., Tanev, H., Kabadjov, M. and Steinberger, R. (2011d): Aspect-Driven News Summarization. In: International Journal of Computational Linguistics and Applications 2 (1-2), Bahri Publications.

Steinberger, J., Kabadjov, M., Steinberger, R., Tanev, H., Turchi, M. and Zavarella, V. (2012a): Towards language-independent news summarization. In: Proceedings of the Text Analysis Conference 2011, National Institute of Standards and Technology (NIST). Gaithersburg, USA.

Steinberger, J. and Turchi M. (2012b): Machine Translation for Multilingual Summary Content Evaluation. In: Proceedings of the NAACL Workshop on Evaluation Metrics and System Comparison for Automatic Summarization, p. 19-27, ACL.

Steinberger. J., Ebrahim, M., Ehrmann, M., Hurriyetoglu, A., Kabadjov, M., Lenkova, P., Steinberger, R., Tanev, H., Vázquez, S. and Zavarella, V. (2012c): Creating sentiment dictionaries via triangulation, In: Decision Support Systems 53, p. 689–694, Elsevier.

Turchi, M., Steinberger, J., Kabadjov, M. and Steinberger, R. (2010): Using Parallel Corpora for Multilingual (Multi-Document) Summarisation Evaluation. In: Multilingual and Multimodal Information Access Evaluation, Springer Lecture Notes for Computer Science 6360, pages 52-63, Springer.

Turchi, M., Atkinson, M., Wilcox, A., Crawley, B., Bucci, S., Steinberger. R. and Van der Goot, E. (2012): Onts:optima news translation system, EACL 2012, p. 25, ACL.

Wan, X., Li, H., Xiao, J. (2010): Cross-language document summarization based on machine translationquality prediction. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 917–926, ACL.

Appendixes

A. Steinberger J., Poesio M., Kabadjov M. and Ježek K.: Two Uses of Anaphora Resolution in Summarization. In: Information Processing & Management 43(6), pages 1663-1680, Elsevier, 2007.

B. Steinberger, J., Belyaeva, J., Crawley, J., Della Rocca, L., Ebrahim, M., Ehrmann, M., Kabadjov, M., Steinberger, R. and van der Goot, E.: Highly Multilingual Coreference Resolution Exploiting a Mature Entity Repository. In: Proceedings of the 8th International Conference Recent Advances in Natural Language Processing (RANLP), pages 254-260, Hissar, Bulgaria, 2011.

C. Kabadjov, M., Steinberger, J. and Steinberger, R.: Multilingual Statistical News Summarization. In: Thierry Poibeau, Horacio Saggion, Jakub Piskorski & Roman Yangarber (eds.), Multi-source, Multilingual Information Extraction and Summarization, pages 229-252, Springer, 2013.

D. Giannakopoulos, G., El-Haj, M., Favre, B., Litvak, M., Steinberger, J. and Varma, V.: TAC 2011 multiling pilot overview. In: Proceedings of the Text Analysis Conference 2011, National Institute of Standards and Technology (NIST), Gaithersburg, USA, NIST, 2011.

E. Turchi, M.., Steinberger, J., Kabadjov, M. and Steinberger, R.: Using Parallel Corpora for Multilingual (Multi-Document) Summarisation Evaluation. In: Multilingual and Multimodal Information Access Evaluation, Springer Lecture Notes for Computer Science 6360, pages 52-63, Springer, 2010.

F. Steinberger, J. and Turchi, M.: Machine Translation for Multilingual Summary Content Evaluation. In: Proceedings of the NAACL Workshop on Evaluation Metrics and System Comparison for Automatic Summarization, pages 19-27, ACL. Montreal, Canada, 2012.

G. Steinberger, J., Ebrahim, M., Ehrmann, M., Hurriyetoglu, A., Kabadjov, M., Lenkova, P., Steinberger, R., Tanev, H., Vázquez, S., Zavarella, V.: Creating sentiment dictionaries via triangulation, In: Decision Support Systems 53, pages 689–694, Elsevier, 2012.

H. Steinberger, J., Lenkova, P., Kabadjov, M., Steinberger, R. and van der Goot, E.: Multilingual Entity-Centred Sentiment Analysis Evaluated by Parallel Corpora. In: Proceedings of the 8th International Conference Recent Advances in Natural Language Processing, pages 770-775. Hissar, Bulgaria, 2011.

I. Balahur, A., Kabadjov, M., Steinberger, J., Steinberger, R. and Montoyo, A.: Challenges and solutions in the opinion summarization of user-generated content. In: Journal of Intelligent Information Systems 39(2), pages 375-398, Springer, 2012.