Aspects of Multilingual News Summarisation

Josef Steinberger
University of West Bohemia, Czech Republic

Ralf Steinberger, Hristo Tanev, Vanni Zavarella
Joint Research Centre, European Commission, Italy

Marco Turchi
Fondazione Bruno Kessler, Trento, Italy

ABSTRACT

In this book chapter, we discuss several pertinent aspects of an automatic system that generates summaries in multiple languages for sets of topic-related news articles (multilingual multi-document summarisation), gathered by news aggregation systems. The discussion follows a framework based on Latent Semantic Analysis (LSA) because LSA was shown to be a high-performing method across many different languages. Starting from a sentence-extractive approach we show how domain-specific aspects can be used and how a compression and paraphrasing method can be plugged in. We also discuss the challenging problem of summarisation evaluation in different languages. In particular, we describe two approaches: the first uses a parallel corpus and the second statistical machine translation.

INTRODUCTION

News gathering and analysis systems, such as *Google News* or the *Europe Media Monitor*[i], gather tens or hundreds of thousands of articles per day. Efforts to summarise such highly redundant news data are motivated by the need to automatically inform news end users of the main contents of up to hundreds of news articles talking on a particular event, e.g. by sending a breaking news text message or an email. Due to the high multilinguality of the raw news data, any summariser must be multilingual.

In this chapter, we first present an overview of summarisation approaches and a discussion of their possible application to other languages. We study deeply one particular approach based on Latent Semantic Analysis (LSA) (Steinberger et al., 2012) because LSA was shown to be a high-performing method across many different languages in the multilingual task of the Text Analysis Conference (TAC[ii]) in 2011. We start from the basic LSA approach (Steinberger and Jezek, 2009).

We then discuss the more challenging task of *aspect-based* summarisation, as defined at TAC'2010[iii]. In the aspect scenario, the goal is to produce a summary from articles about a specific event which falls into a predefined domain (e.g. terrorist attacks), for which we have defined aspects that should be mentioned in the summary (e.g. what, when, where happened; who were the victims, perpetrators, etc.). This scenario forces systems to make use of information extraction and to look at the content selection from a more semantic point of view. We will show how an event extraction system can be used to detect pieces of required information and then to extract the related content (Steinberger et al., 2011).

The majority of approaches to automatically summarising documents are limited to selecting the most important sentences. We will therefore dedicate some effort to discussing sentence compression/paraphrasing approaches aiming at more human-like summaries, which typically consist of shorter sentences than automatic summaries. As the ultimate goal is to apply the approach to multiple languages, we will discuss how far we can get with a statistical sentence compression/paraphrasing method (Steinberger et al., 2010).

TAC/DUC evaluation campaigns were the most important events to perform large-scale experiments and discuss evaluation methodology in the last years. We follow the TAC roadmap and discuss the multilingual issue. Evaluation of automatically produced summaries in different languages is a challenging problem for the summarisation community because human efforts are multiplied to create model summaries for each language. At TAC'11, six research groups spent a considerable effort 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 language, write model summaries and evaluate the peer summaries for all the languages. We will discuss findings of the TAC's multilingual task which was the first shared task to evaluate summaries in more than two languages. We will then propose two possibilities how to lower the huge annotation costs:

First, we will consider using a parallel corpus for the multilingual evaluation task. Because of the unavailability of parallel corpora suitable for news summarisation we will follow an effort to create such a corpus (Turchi et al., 2010). The approach is based on the manual selection of the most important sentences in a cluster of documents from a sentence-aligned parallel corpus, and by projecting the sentence selection in one language to various target languages. Although model summaries were not created, and texts were taken from a slightly different genre (news commentaries), the evaluation results are directly comparable across languages.

Second, we will discuss using Machine Translation (MT) to achieve multilingual summarisation evaluation. In the last fifteen years, research on 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 can translate text into more than 50 languages, showing that MT is not a pipe-dream. We thus investigate how machine translation can be plugged in to evaluate summaries in other languages. We will try to see whether machine-translated models can perform close to manually created evaluation models (Steinberger and Turchi, 2012).

The remainder of the chapter is organized as follows: It contains two main sections, each discussing several aspects: In the first section, we describe several approaches aimed at building coherent summaries by selecting the most informative sentences from a set of documents. We also describe the specific case of aspect-based summarisation, importance of temporal analysis and that of compression and paraphrasing techniques. The second section is dedicated to issues concerning the evaluation of multilingual summaries. With that regard, we will investigate how parallel data and how statistical machine translation can be used. After the core two sections, we follow performance of the LSA-based summariser, as a representative of multilingual approaches, at the TAC 2008-2011 evaluations in the results section.

BACKGROUND

Automatic news summarisation deals with the problem of producing a succinct informative gist for a set of news articles about the same topic. The aim of the task could be that the target language of the summary be the same as the input articles (standard single-/multi-document summarisation) (Nenkova and Louis, 2008) or that the languages of summary/input articles be different (cross-language document summarisation) (Wan et al., 2010). Moreover, the task of handling several languages, with summary and input articles being in the same language, has been termed as multilingual summarisation (Litvak et al., 2010).

Summarisation has been an active area of research for several decades (Luhn, 1958; Edmundson, 1969), but in particular over the past seventeen years. The area initially focused on single-document summarisation (Mani and Maybury, 1999), a fact reflected by the first US NIST-organized *Document Understanding Conference* (DUC) evaluation exercises (Over et al., 2007). Then, over the past decade the emphasis shifted to multi-document summarisation exemplified by latter DUCs followed by the *Text Analysis Conference* (TAC). However, it has been only recently that interest in multilingual summarisation has risen (Kabadjov et al., 2013; Litvak et al., 2010).

MULTILINGUAL SUMMARISATION

Extractive summarisation

Work on Text Summarisation has been quite varied and abundant. A basic processing model for Text Summarisation, proposed by Sparck-Jones (1999) comprises three main stages: source text interpretation (I) to construct a source representation (e.g., lexical chains, semantic graphs, discourse models), source representation transformation (T) to form a summary representation (e.g., Singular Value Decomposition, SVD), and summary text generation (G). More practically-motivated approaches that use shallow linguistic analysis and only partially cover this processing model, as well as more ambitious ones attempting all three stages using deep semantic analysis have been proposed in the literature.

There are approaches based on shallow linguistic analysis such as word frequencies (Luhn, 1958), cue phrases (e.g., “in conclusion”, “in summary”) and location (e.g., title, section headings) (Edmundson,

1969); there are machine learning approaches that combine a number of surface features (Kupiec et al., 1995) and/or more elaborate features exploiting discourse structure (Teufel and Moens, 1999) to train classifiers using specialized corpora formed by pairs of documents and their hand-written summaries; there are also more sophisticated approaches, but still working at the surface level, exploiting cohesive relations like co-reference (Boguraev and Kennedy, 1999) and lexical cohesion (Barzilay and Elhadad, 1999) to identify salience or purely lexical approaches trying to identify 'implicit topics' by conflating words using methods inspired by Latent Semantic Analysis (LSA) (Gong and Liu, 2002); using yet deeper linguistic analysis, there are approaches purely based on discourse structure (e.g., RST) (Marcu, 1999) and others combining discourse structure with surface features (Hovy and Lin, 1999) or lexical with higher level semantic information such as anaphora (Steinberger et al., 2007); and finally there are knowledge-rich approaches, where the source undergoes a substantial semantic analysis during the process of filling in a predefined template (McKeown and Radev, 1995) or the source data is available in a more structured way (i.e., events have been identified already) (Maybury, 1999).

Summarisation based on LSA

Approaches based on term co-occurrence (e.g. LSA) represent a good base for building a language-independent (or multilingual) summariser. The LSA approach (Steinberger and Ježek, 2009) first builds a term-by-sentence matrix from the source, then applies Singular Value Decomposition (SVD) and finally uses the resulting matrices to identify and extract the most salient sentences. SVD finds the latent (orthogonal) dimensions, which in simple terms correspond to the different topics discussed in the source.

More formally, it constructs the terms by sentences association matrix $\boldsymbol{A}$. Each element of $\boldsymbol{A}$ indicates the weighted frequency of a given term in a given sentence. Having m distinguished terms and n sentences in the documents under consideration the size of $\boldsymbol{A}$ is m x n. Element a_{ij} of $\boldsymbol{A}$ represents the weighted frequency of term i in sentence j and is defined as:

$$a_{i,j} = L_{i,j} \cdot G_i,$$

where $L_{i,j}$ is the local weight of term i in sentence j and G_i is the global weight of term i in the document set. The weighting scheme found to work best (Steinberger et al., 2007) uses a binary local weight and an entropy-based global weight:

$$L_{i,j} = 1 \text{ if term } i \text{ appears at least once in sentence } j\text{; otherwise } L_{i,j} = 0,$$

$$G_i = 1 - \sum_{j=0}^{j<n} \frac{p_{i,j} \log p_{i,j}}{\log n}, p_{i,j} = \frac{t_{i,j}}{g_i},$$

where $t_{i,j}$ is the frequency of term i in sentence j, g_i is the total number of times that term i occurs in the whole set of documents and n is the number of sentences in the set.

After that step Singular Value Decomposition (SVD) is applied to the above matrix. The SVD of an *m* x *n* matrix is defined as:

$$A = U \cdot S \cdot V^T,$$

where $\boldsymbol{U}$ (*m* x *n*) is a column-orthonormal matrix, whose columns are called left singular vectors. The matrix contains representations of terms expressed in the newly created (latent) dimensions. $\boldsymbol{S}$ (*n* x *n*) is a diagonal matrix, whose diagonal elements are non-negative singular values sorted in descending order. $\boldsymbol{V}^{\mathrm{T}}$ (*n* x *n*) is a row-orthonormal matrix which contains representations of sentences expressed in the latent dimensions. The dimensionality of the matrices is reduced to *r* most important dimensions and thus, we receive matrices $\boldsymbol{U}'$ (*m* x *r*), $\boldsymbol{S}'$ (*r* x *r*) a $\boldsymbol{V}'^{\mathrm{T}}$ (*r* x *n*). The value of *r* can be set according to the summarisation ratio (*r = summarisation_ratio · n).* For example, having 200 sentences to be summarised to 5%, *r* will be set up to 10. Another possibility is to learn the optimal *r* from the training data (Steinberger and Ježek, 2009).

From the mathematical point of view SVD maps the *m*-dimensional space specified by matrix $\boldsymbol{A}$ to the *r*-dimensional singular space. From an NLP perspective, what SVD does is to derive the latent semantic structure of the document set represented by matrix $\boldsymbol{A}$: i.e. a breakdown of the original documents into *r* linearly-independent base vectors which express the main 'topics' of the document set. SVD can capture interrelationships among terms, so that terms and sentences can be clustered on a 'semantic' basis rather than on the basis of words only. Furthermore, as demonstrated in [4], if a word combination pattern is salient and recurring in a document set, this pattern will be captured and represented by one of the singular vectors. The magnitude of the corresponding singular value indicates the importance degree of this pattern within the document set. Any sentences containing this word combination pattern will be projected along this singular vector, and the sentence that best represents this pattern will have the largest index value with this vector. Assuming that each particular word combination pattern describes a certain topic in the document (LSA topic), each singular vector can be viewed as representing such a topic (Ding, 2005), the magnitude of its singular value representing the degree of importance of this topic.

Matrix $\boldsymbol{V^T}$ contains representation of sentences in the LSA topics and $\boldsymbol{S}$ contains importance of those topics. Thus their product, matrix $\boldsymbol{F = S \cdot V^T}$, represent the sentence latent space weighted by topic importance. Sentence selection starts with measuring the length (Euclidean norm) of sentence vectors in matrix $\boldsymbol{F}$. The length of the vector can be viewed as a measure for importance of that sentence within the top LSA topics. It was called *co-occurrence sentence score* in (Steinberger et al., 2011). The sentence with the largest score is selected as the first to go to the summary (its corresponding vector in $\boldsymbol{F}$ is denoted as $\boldsymbol{f}_{best}$). After placing it in the summary, the topic/sentence distribution in matrix $\boldsymbol{F}$ is changed by subtracting the information contained in that sentence:

$$\boldsymbol{F}^{(it+1)} = \boldsymbol{F}^{(it)} - \frac{\boldsymbol{f}_{best} \cdot \boldsymbol{f}_{best}^T}{\|\boldsymbol{f}_{best}\|^2} \cdot \boldsymbol{F}^{(it)} \qquad (1)$$

The vector lengths of similar sentences are decreased, thus preventing within-summary redundancy. After the subtraction of information in the selected sentence, the process continues with the sentence which has the largest co-occurrence sentence score computed on the updated matrix $\boldsymbol{F}$. The process is iteratively repeated until the required summary length is reached.

Aspect-driven summarisation

TAC'10 encouraged a deeper semantic analysis of the source documents by its new Guided summarisation task. Summarisers were given a list of aspects for each article category, and the summary should include those aspects if possible. The task naturally led to the integration with information extraction tools.

Steinberger et al. (2011) proposed an approach which works in multiple languages. It used an event extraction system that is focused on similar issues as the categories defined for TAC'10. For capturing other aspects they automatically learned terms semantically related to a manually created set of seed terms. The aim was to select frequently mentioned information, whilst at the same time making sure that this information also captures the required aspects. Thus, a combination of the co-occurrence-based information from LSA and the aspect information coming from the event extraction system was proposed.

The event extraction system (NEXUS), which was used in the experiments, analyses news articles reporting on violent events, natural or man-made disasters (see Tanev et al. (2008) for a detailed system description). The system identifies the type of the event (e.g., flooding, explosion, assassination, kidnapping, air attack, etc.), number and description of the victims, as well as descriptions of the perpetrators and the means used by them. For example for the text "*Three people were shot dead and five were injured in a shootout*", NEXUS will return an event structure with three slots filled: The *event type slot* will be set to *shooting*; the *dead victims slot* will be set to *three people*; and the *injured slot* will be set to *five*. Event extraction is deployed as a part of the EMM family of applications, described in (R.Steinberger et al., 2009). NEXUS relies on a mixture of manually created linguistic rules, linear patterns, acquired through machine learning procedures, plus domain knowledge, represented as domain-specific heuristics and taxonomies. In the summarisation experiments the event extraction system was run on each news article from the corpus and the extracted slots were mapped to the summarisation aspects.

It was found out that some of the aspects, relevant to the summarisation task, correspond to the information extracted by NEXUS. In particular, the aspects *what happened, perpetrators* and *who affected* have corresponding slots in the event structures of NEXUS.

In our summarisation experiments we ran the event extraction system on each news article from the corpus and we mapped extracted slots to summarisation aspects. This was done in the following way: The event type (e.g., terrorist attack) was mapped to the aspect *what happened*; the slot *perpetrator* was mapped to the aspect *perpetrators*; and the values for the aspect *who affected* were obtained as a union of the event slots: *dead victims*, *injured*, *arrested*, *displaced*, *kidnapped*, *released hostages* and *people left without homes*. At the end, from a fragment like: "*three people died and many were injured*", the system will extract two values for the aspect *who affected*, namely "*three people*" and "*many*".

For the other aspects lexica were generated using *Ontopopulis*, a system for the automatic learning of semantic classes, based on distributional semantics (see Tanev et al. (2006) for algorithm overview and evaluation). As an input, it accepts a list of words, which belong to a certain semantic class, e.g. "*disasters*", and then it learns additional words, which belong to the same class, e.g. *"earthquake"*, *"flooding"*, etc. Clearly, the system output needs to be manually cleaned, in order to build an accurate

lexicon. Since the terms are ordered by reliability (more reliable terms are at the top), the users can review the list top-down deciding where to stop on the basis of their availability or the quality of the list around the point reached within the list. The unrevised items are discarded. Another possibility is to skip the manual reviewing process and take all the terms up to a certain threshold. This approach, however, cannot guarantee very high accuracy.

Four lexica were learned using Ontopopulis, followed by manual cleaning. Each lexicon was relevant to a specific summary aspect. The four aspects covered by our lexica are: "*damages*", "*countermeasures*", "*resource*", and "*charges*". Here we give a short sample from each of the learned lexica:

1. damages: damaged, destroyed, badly damaged, extensively damaged, gutted, torched, severely damaged, burnt, burned;
2. countermeasures: operation, rescue operation, rescue, evacuation, treatment, assistance, relief, military operation, police operation, security operation, aid;
3. resource: water, food, species, drinking water, electricity, gas, forests, fuel, natural gas;
4. charges: rape, kidnapping, aggravated, murder, attempted murder, robbery, aggravated assault, theft, armed robbery.

The words and multi-word units from these four lexicons were used to trigger the corresponding summary aspects.

The identified aspects were used to boost the co-occurrence-based scores of the sentences that contained them. For each document set an aspect-by-sentence matrix which contained Boolean values to store an aspect's presence/absence in sentences was built. The length of the sentence vector in the aspect matrix worked as a booster for the co-occurrence score. After selecting a sentence, the influence of the aspects mentioned there was lowered. For details see Steinberger et al. (2011).

Temporal analysis

Temporal analysis is important in various summarisation subtasks. It can help to identify date and time of the event in the case of event-focused topics (the *when* aspect in the case of the guided summarisation task). It can provide important features for summary sentence ordering in the case of story-focused summaries. Steinberger et al. (2012) integrated into the processing chain a temporal analysis module which deals with the detection and normalisation of so called temporal expressions (timex), whose extent and classification are defined in partial compliance with the TIMEX2 standard (Ferro et al., 2005).

Temporal expressions are categorized into a small set of temporal types (Date, Duration, Period and Set), and include constructions such as numerical and non-numerical dates, underspecified dates ("in March 2002"), absolute, relative or deictic expressions (e.g. "in March 2002", "in March", "last month", respectively), fuzzy time references ("in a few months") and their compositions ("a year before last Monday").

Timex processing consists of a Recognition and a Normalization stage. In the Recognition phase, they are detected and segmented in text through local finite-state parsing performed by a cascade of hand-coded, partially language-dependent rules. Rules build a more abstract, intermediate typed feature structure-like representation of the temporal expressions, which is then exploited by a language-independent Normalization module. This latter performs, first, "anchor selection", that is it determines and maintains a reference time for relative timex resolution, by starting using the article creation date and updating it along the resolution process according to a simple heuristic: find the closest preceding resolved timex, within the same sentence, with a compatible level of granularity (e.g. a relative timex such as "in the afternoon" can be resolved by a day granularity timex like "the day after" but not by "in 2010"). Then, it uses the reference time to resolve relative timexes, computes exact calendar values of the time expressions and finally normalizes them according to the machine-readable TIMEX2 standard (Zavarella and Tanev, in press) Finally, the most frequent normalized timex of type Date in the article set was simply taken as the time of the target event. The whole method is highly limited in recall by exclusively focusing on explicit temporal expressions and ignoring other temporal relation markers. Also, anaphoric event references in text are not detected, so that a sentence containing, for example, the phrase "*after the attack on Bagdad*", will not trigger any additional piece of temporal information, with respect to a previously detected terrorist attack event.

Compression and paraphrasing techniques

Empirical evidence shows that human summaries contain on average more and shorter sentences than system summaries (6 sentences vs. 4 sentences in TAC'09 data – Turchi et al., 2010). By compressing and/or rephrasing, the saved space in the summary could be filled in by the next most salient sentences, and thus the summary can cover more content from the source texts. Turchi et al. (2010) and Steinberger et al. (2012) tried to investigate language-independent possibilities to achieve this goal. The initial experimental results showed that the approach is feasible, since it produced summaries, which when evaluated against the TAC'09 data yield ROUGE scores comparable to the average of the participating systems. However, it achieved lower scores compared to the sentence-extractive summariser.

The approach starts by identifying the most salient terms in each selected sentence. For each term it computes the term salience score from the LSA. Matrix $\boldsymbol{U}$ contains representation of terms in the LSA topics and $\boldsymbol{S}$ contains importance of those topics. Thus their product, matrix $\boldsymbol{T} = \boldsymbol{U} \cdot \boldsymbol{S}$, represents the term latent space weighted by topic importance. Notice the analogy with the LSA-based sentence selection approach. For each term *i*, the salience score is given by $\|\mathbf{t}_i\|$. In addition, language model probabilities up to 4-grams were computed. The salience score should reflect the local importance of the term within the document set (mainly nouns) and language model probabilities should add the globally important terms (e.g. verbs). After normalising scores of each feature and combining them, each term ended up with a score reflecting its importance in the sentence. The final term sequence consisted of the top 70% terms. To make the sequence more readable, the sentences were reconstructed by the noisy-channel model primarily used by SMT systems, adding the most probable content (mainly stopwords) to connect the sentence fragments. The interpretation of the noisy channel in this application consists of looking at a stemmed string without stopwords and imagining that it was originally a long string and that someone

removed or stemmed some text from it. In the proposed framework, reconstruction consists of identifying the original long string (for details see Turchi et al. (2010)). The term selection gives compression capabilities and the reconstruction adds paraphrasing capabilities.

SUMMARY EVALUATION IN MULTIPLE LANGUAGES

Evaluation of automatically produced summaries in different languages is a challenging problem for the summarisation community, because human efforts are multiplied to create model summaries for each language. Unavailability of parallel corpora suitable for news summarisation adds even another annotation load because documents need to be translated to other languages. At the last TAC'11 campaign, six research groups created evaluation resources in seven languages (Giannakopoulos et al., 2012). 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 language, write model summaries and evaluate the peer summaries for all the languages.

The Multilingual task of TAC'11 (Giannakopoulos et al., 2012) aimed at evaluating the application of (partially or fully) language-independent summarisation algorithms on a variety of languages. The task was to generate a representative summary (250 words) of a set of 10 related news articles. It included 7 languages (English, Czech, French, Hebrew, Hindi, Greek and Arabic). Annotation of each language sub-corpus was performed by a different group. English articles were manually translated to the target languages, 3 model summaries were written for each topic. Eight groups (systems) participated in the task; however, not all systems produced summaries for all languages. Human annotators scored each summary, both models and peers, on a 5-to-1 scale. The score corresponded to the overall responsiveness of the main TAC task – equal weight of content and readability.

Although the manually assigned grades showed a clear gap between human and automatic summaries, there were 5 systems for English and 1 system for French which were not significantly worse than at least one model. ROUGE (Lin, 2004) is widely used for English evaluations because of its simplicity and its high correlation with manually assigned content quality scores on overall system rankings, although per-case correlation is lower. As it compares n-grams in a system and reference summaries, it is possible to use it for evaluating summaries in other languages. However, it performs worse in the case of specific languages (e.g. languages with free word order) and more effort should be allocated to find a more appropriate set of evaluation methods. Although using n-grams with n greater than 1 gives limited possibility to reflect readability in the scores when compared to reference summaries, ROUGE is considered mainly as a content evaluation metric.

Using parallel data

A method, and related resources, which allows saving precious annotation time and that makes the evaluation results across languages directly comparable was introduced by Turchi et al. (2010). This approach relies on parallel data and it is based on the manual selection of the most important sentences in a cluster of documents from a sentence-aligned parallel corpus, and by projecting the sentence selection to various target languages.

In extractive summarisation, a single or multi-document summary is produced by selecting the most relevant sentences. It can then be evaluated by comparing these sentences with a gold standard of manually selected sentences. If sentence alignment information is available for a parallel text collection, the gold standard of one language can be projected to all the other languages. The more languages there are in the parallel corpus, the more time can be saved with this method. Sentences are not always aligned one-to-one because a translator may decide, for stylistic or other reasons, to split a sentence into two or to combine two sentences into one. Translations and original texts are never perfect, so that it is also possible that the translator accidentally omits or adds some information, or even a whole sentence. For these reasons, aligners such as Vanilla[vi], which implements the Gale and Church (2012) algorithm, can be used.

Using statistical machine translation

Steinberger and Turchi (2012) addressed the same problem of reducing annotation time and generating models, but from a different perspective. Instead of using parallel data and annotation projection or full documents, they investigated the use of machine translation at a different level of summary evaluation. While the approach of Turchi et al. (2010) is focused on sentence selection evaluation, the strategy of Steinberger and Turchi (2012) can also evaluate generative summaries, because it works at the summary level.

When we want to evaluate a summary on language A and we have evaluation resources in language B we can translate the summary to language A. Another approach, investigated by Steinberger and Turchi (2012), is to produce models in one pivot language (e.g., English) and translate them automatically to all other languages. In the results section we discuss the results of using both machine-translated model and system summaries.

RESULTS

In this section, we will show how the language-independent LSA-based summariser performs on different datasets. The next part shows how the summariser can compete with systems adapted for English at the TAC evaluation campaigns. A manual multilingual evaluation will then show how the summariser behaves when it is run on various languages. And finally, the two approaches, which lower the annotation costs, will be evaluated: using parallel and translated data.

LSA-based summariser vs. state-of-the-art on English (TAC 2008-2011)

In this section we report the results of the multilingual LSA-based summariser in all four TAC evaluation exercises. In 2008 and 2009, the basic summarisation task required systems to produce a short (100 words) summary of a set of English newswire articles (initial summaries). The update task aimed at producing a summary of a set of chronologically newer articles under the assumption that the user has

already read a given set of earlier articles (used for the creation of an initial summary). In 2010 and 2011, the automatic summaries were supposed to include predefined category-related aspects resulting in the Guided summarisation task. The summaries submitted by participating systems were evaluated against human-created summaries based on various evaluation measures. Overall responsiveness evaluated the degree to which a summary is responding to the information need contained in the topic statement, considering the summary's content as well as its linguistic quality. In Table 1, we report and compare overall responsiveness scores of different variants of the multilingual LSA-based summariser.

	Best system	LSA	+ entities	+ aspects	+ temporal analysis	+ compression
TAC 2008	2.792	2.667 (10/58)				
TAC 2009	3.080		2.978 (2/52)			
TAC 2010	3.170		2.890 (19/43)	2.980 (10/43)		
TAC 2011	3.159				2.977 (12/50)	2.341 (43/50)

Table 1: Overall responsiveness score of the multilingual LSA-based summariser and its variants throughout the TAC 2008 – 2011 evaluations (English initial summaries). It shows the best TAC system's score. LSA = system based on latent semantic analysis, +entities = addition of entities into the LSA matrix in 2009 was discussed in Kabadjov et al.(2013), +aspects = sentence selection also based on the aspect information, +compression = compression/paraphrasing method included. Score (rank/total number of participating systems). The scale is from 5=very good to 1=very poor.

In 2008, the summaries of the LSA-based summariser were ranked 10th from a total of 58 participating systems. They were not statistically significantly worse than those of the best TAC system indicating that even a simple language-independent summariser can perform close to the best systems for English. In 2009, the system with more semantic representation, which included entities (see Kabadjov et al. (2013)), was ranked 2nd overall although the approach was multilingual. The TAC 2010 and TAC 2011 LSA-based summariser, which extracted and used aspects for the guided summarisation task, was ranked still at the state-of-the-art level (10th out of 43 and 12th out of 50). The sentence-generative summaries submitted to TAC 2011 suffered from worse readability, which also affected content-based scores. This showed that they still cannot compete with sentence-extractive summaries.

Mutlilingual manual evaluation (TAC Mutliling 2011)

TAC 2011 included a pilot multilingual evaluation task. The aim was to generate again a representative summary of a set of 10 documents describing an event sequence – a set of atomic event descriptions, sequenced in time, that share main actors. An important difference compared to the main TAC summarisation task was that the limit of summary length was set to 250 words. Human annotators scored

each summary on the 5-to-1 scale, similarly to the overall responsiveness of the main TAC task, with equal weight of content and readability.

The LSA-based system received the highest score in the case of 5 languages – Czech, English, French, Hebrew and Greek (see Table 2). For Arabic it was lower than baseline (the start of the centroid article) and for Hindi three other systems performed better. Looking at the average across languages, the summariser received a score of 3.37 indicating positive above-average (> 3) quality of the produced summaries. The basic lexical version of the summariser was used for the experiments. The only resource dependent on the language was a list of stopwords. It did not use the entity detection, event extraction and temporal analysis tools because they had not been developed for all the languages of the task yet. However, all these extensions were designed to work highly multilingually. So far, NER has been introduced for 20 languages, event extraction for 8 languages and temporal analysis for 4 languages. Thus these extensions could improve the summariser's performance at the cost of limiting the set of target languages.

	Best system	LSA
Arabic	3.77	3.43 (4/9)
Czech	3.40	3.40 (1/7)
English	3.57	3.57 (1/10)
French	3.23	3.23 (1/9)
Hebrew	3.87	3.87 (1/7)
Hindi	2.73	2.47 (4/9)
Greek	3.63	3.63 (1/7)

Table 2: Overall responsiveness score of the TAC 2011 Mutliling task for all languages. The best system's score and the one of the LSA-based are reported. Score (rank/total number of participating systems).The scale is from 5=very good to 1=very poor.

Evaluation on parallel data

This section will discuss the evaluation on a parallel corpus done by Turchi et al. (2010). The binary model considered a sentence important if it was selected by at least two annotators (there were 4 annotations in total). The score of a system summary can then be computed as: the number of sentences in the intersection between the system summary and the sentences selected by at least two annotators divided by the number of sentences in the system summary. Table 3 shows the results of this evaluation approach.

	Random	Lead	LSA
Arabic	22%	25%	60%
Czech	21%	25%	70%
English	21%	25%	60%
French	21%	25%	45%
German	22%	20%	55%
Russian	24%	25%	50%
Spanish	21%	30%	50%

Table 3: Summary evaluation on the parallel corpus from Turchi et al. (2010) using binary model - summary length is 10 sentences. It compares the LSA approach with 2 baselines: random sentence selection (Random) and selecting the first sentences from each article (Lead). The values are ratios of the number of sentences in the intersection between the system summary and the sentences selected by at least two annotators divided by the number of sentences in the system summary (10).

We can see that the LSA summarizer selected on average over 5 sentences from 10 (56%) that at least two annotators marked as important. We can also observe that it is significantly better than baselines for all languages. It was also interesting to see that such a language-independent summariser selects on average only 35% of the same sentences for a language pair. Agreement peaks do exist, like the Czech-Russian pair (41%), which may be due to the fact that they are both Slavic languages and thus have similar properties. This indicates that there is a real need for multilingual summarisation evaluation, even if the summariser in principle uses only statistical, language-independent features.

Evaluation on translated data

The study in Steinberger and Turchi (2012) addressed the same problem. It investigated whether we can annotate in one language and instead of using sentence selection and its projection to other languages in a parallel corpus they proposed to use Machine translation. This approach is not constrained on sentence selection as the one using parallel corpus and thus the experiments were run on the TAC Multiling 2011 corpus.

There are two basic possibilities how we can evaluate a summary in language B given that we have model summaries in language A. Table 4 compares translating a system summary from language B to A and translating models from A to B. Thanks to the nature of the multilingual corpus, we can use translation of models from more languages. We analyse the ROUGE-SU4 score (Lin, 2004), which is more suitable than ROUGE-2 for languages with free word order. The figures show correlations of system rankings (both model and system summaries included) provided by ROUGE to system rankings generated using grades manually assigned to each summary.

Three languages were analysed: English, French and Czech. The translation approach discussed in (Turchi et al., 2012) was used to build four models covering the following language pairs: En-Fr, En-Cz, Fr-En and Cz-En. Performance was evaluated using the Bleu score (Papineni et al., 2002): En-Fr 0.23, En-Cz 0.14, Fr-En 0.26 and Cz-En 0.22.

Evaluation language	Model summaries	System summaries	Translation quality (BLEU)	Avg. system ROUGE-SU4	R.-SU4&grades correlation
EN	EN models	EN summaries		.183	.723 (p<.02)
	EN models	FR summaries translated to EN	0.26	.184	.581 (p<.05)
	EN models	CZ summaries translated to EN	0.22	.217	.777 (p<.02)
	FR models translated to EN	EN summaries	0.26	.170	.785 (p<.01)
	CZ models translated to EN	EN summaries	0.22	.162	.692 (p<.02)
	FR&CZ models translated to EN	EN summaries	0.26 & 0.22	.153	.759 (p<.01)
FR	FR models	FR summaries		.207	.700 (p<.02)
	FR models	EN summaries translated to FR	0.23	.190	.839 (p<.01)
	EN models translated to FR	FR summaries	0.23	.186	.559 (p<.1)
CZ	CZ models	CZ summaries		.211	.636 (p<.1)
	CZ models	EN summaries translated to CZ	0.14	.160	.620 (p<.05)
	EN models translated to CZ	CZ summaries	0.14	.172	.651 (p<.1)

Table 4: Evaluation of using machine-translated data from Steinberger et al. (2012). Each row corresponds to one evaluation settings: using original/translated models (2nd column), evaluating original/translated system summaries (3rd column), Bleu score capturing quality of the machine translator, an average of ROUGE-SU4 system scores and the correlations of the system ranking provided by ROUGE to the system ranking based on manually-assigned grades.

The results indicate (correlations were not always enough statistically significant, see the p-value) that the use of translated models or summaries did not alter much the overall system ranking. A drop in ROUGE score was evident, and it strongly depended on the translation performance.

FUTURE RESEARCH DIRECTIONS

We have seen that our automatic summariser does not select the same sentences in the different language versions of a parallel corpus. As this behaviour is relevant for the language independence assumption of our system, we feel that we need to analyse the reasons for this different sentence selection.

Regarding the selection of the most important sentences by the human annotators, we would like to deepen our insights regarding the human choice of summary-worthy sentences. Besides the fact that the most relevant aspects should be covered in a summary, it may be necessary to look at the human selection behaviour in enumerations, such as pros and cons regarding a certain argument or subject.

In our work up to now, we have identified a whole range of issues that would improve the summary result. We aim at improving each of them.

Regarding aspect-driven summarisation, we are interested in automatically identifying for any news cluster what the most relevant and important expected information aspects are and to focus the summary on those. Identifying these aspects could be achieved through a rough classification of the news and its

vocabulary into major event classes. Identifying the *When*, *Where* and *Why* aspects is challenging, especially for multiple languages.

Regarding compression and paraphrasing, we are considering using more features and working harder on making the summaries more readable.

A thorough human analysis of the current results of our basic sentence selection summaries shows that using some simple heuristics will increase the readability a lot and it will at the same time reduce the summary length, leaving more space for further information aspects and details. Steps we currently work on are to use a word n-gram overlap measure to reduce the redundancy and repetition in sentences and phrases, such as the full titles of entities mentioned. Another promising step is to delete or avoid location and source information frequently found in first sentences of news articles as these are misleading and also disturbing when found in the middle of a summary.

Regarding summary evaluation, we plan to participate in the *Multiling* workshop at ACL'2013, by adding more languages and extending the evaluation corpus. Running the Machine Translation evaluation experiments on such an extended corpus will help us find more evidence. We would also be interested in researching evaluation methods that are more meaningful than ROUGE and that would work with different languages.

The study of Wan et al. (2011) investigated the task of finding and summarising the major differences between the news articles about the same event in two languages. Also our ultimate goal is to produce a real cross-lingual summariser that is able to identify new information aspects across languages in order to benefit from information complementarity across languages. Incorporating a sentiment analysis component would enable to analyse also different opinions about the same event expressed in different languages.

CONCLUSION

Automatic multi-document summarisation is an important task which has the potential to reduce information overload and to enable users finding more information in less time. The task is rather challenging and complex. Many research avenues can be explored to build summarisers and to improve them. For highly multilingual summarisation software, the methods must be kept relatively simple (e.g. by mostly using statistics, machine learning and annotation projection) as any extended human effort is prohibitive for most system developers. The performance ranking at the multilingual summary evaluation task at TAC2011 of the LSA and entity-based system showed that – with enough effort and resources – one can do better, but having achieved the top performance in most of the non-English languages also showed that this simple method produces results that can be considered a rather high-level baseline that is not that easy to pass. Working on paraphrasing and on sentence compression allows interesting and promising extensions to the base sentence selection approach. The two strands of studies carried out on evaluating automatic summaries in various languages while keeping the human annotation effort low – using annotation projection in parallel corpora and making use of machine translation – help close the annotation effort gap so that such solutions are quintessential for anybody working on summarisation covering several languages.

REFERENCES

Barzilay, R. & Elhadad, M. (1999). Using lexical chains for text summarization. In *Advances in Automatic Text Summarization*. MIT Press.

Boguraev, B. & Kennedy, C. (1999). Salience-based content characterisation of text documents. In *Advances in Automatic Text Summarization*. MIT Press.

Ding C. (2005). A probabilistic model for latent semantic indexing. In *Journal of the American Society for Information Science and Technology 56(6)*.

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

Ferro, L., Gerber, L., Mani, I., Sundheim, B. & Wilson, G. (2005). *TIDES 2005 Standard for the Annotation of Temporal Expressions. Technical Report*, The MITRE Corporation.

Giannakopoulos, G., El-Haj, M., Favre, B., Litvak, M., Steinberger, J. & Varma, V. (2012). TAC 2011 Multiling pilot overview. In *Proceedings of the Text Analysis Conference 2011*, NIST.

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

Hovy, E., Lin, C. (1999). Automated text summarization in Summarist. In *Advances in Automatic Text Summarization*. MIT Press.

Jones, K.S. (1999). Automatic summarising: Factors and directions. In *Advances in Automatic Text Summarization*. MIT Press.

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)*. IEEE/ACM.

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

Kupiec, J., Pedersen, J. & Chen, F. (1995). A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 68–73). ACM.

Lin, C.Y. (2004). ROUGE: a package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*. 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* (pp. 927–936). ACL.

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

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

Marcu, D. (1999). From discourse structures to text summaries. In: *Advances in Automatic Text Summarization*. MIT Press.

Maybury, M. (1999). Generating summaries from event data. In: *Advances in Automatic Text Summarization*. MIT Press.

McKeown, K. & Radev, D. (1995). Generating summaries of multiple news articles. In: *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 74–82). ACM.

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* (pp. 825–833). ACL.

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

Pouliquen, B., Kimler, M., Steinberger, R., Ignat, C., Oellinger, T., Blackler, K., Fuart, F., Zaghouani, W., Widiger, A., Forslund, A.C. & Best, C. (2006). Geocoding multilingual texts: Recognition, disambiguation and visualisation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 53–58). ELRA.

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

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

Steinberger, J. & Ježek, K. (2009) Update summarization based on novel topic distribution. In *Proceedings of the 9th ACM DocEng Conference*, ACM.

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

Steinberger, J., Kabadjov, M., Pouliquen, B., Steinberger, R. & Poesio, M. (2010b). WB-JRC-UT's Participation in TAC 2009: Update Summarization and AESOP Tasks. In *Proceedings of the Text Analysis Conference 2009*, NIST.

Steinberger, J., Tanev, H., Kabadjov, M. and Steinberger, R. (2011). 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. & Zavarella, V. (2012). Towards language-independent news summarization. In *Proceedings of the Text Analysis Conference 2011*, NIST.

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

Steinberger, R., Pouliquen, B. & van der Goot, E. (2009). An introduction to the europe media monitor family of applications. In *Information Access in a Multilingual World Proceedings of the SIGIR*, ACM.

Tanev, H. & Magnini, B. (2006). Weakly supervised approaches for ontology population. In *Proceedings of the 11th conference of the European Chapter of the Association for Computational Linguistics (EACL)*. ACL.

Tanev, H., Piskorski, J., & Atkinson, M. (2008). Real-time news event extraction for global crisis monitoring. In *Proceedings of 13th International Conference on Applications of Natural Language to Information Systems (NLDB 2008)*.

Turchi, M., Steinberger, J., Kabadjov, M. & Steinberger, R. (2010). Using Parallel Corpora for Multilingual (Multi-Document) Summarisation Evaluation. In: *Multilingual and Multimodal Information Access Evaluation, LNCS 6360* (pp. 52-63), Springer.

Turchi, M., Atkinson, M., Wilcox. A., Crawley, B., Bucci. S., Steinberger, R. & van der Goot, E. (2012). ONTS: optima news translation system. In *Proceedings of EACL 2012* (pp. 25-32). ACL

Teufel, S. & Moens, M. (1999). Sentence extraction as a classification task. In: *Advances in Automatic Text Summarization.* MIT Press.

Wan, X., Jia, H., Huang, S., Xiao, J. (2011). Summarizing the differences in multilingual news. In *Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval* (pp. 735–744). ACM.

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

Zavarella V. & Tanev H. (2013). FSS-TimEx for TempEval-3: Extracting Temporal Information from Text. Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)(pp.58-63).ACL.

ADDITIONAL READING SECTION

Extractive summarisation:

Jones, K.S. (1999). Automatic summarising: Factors and directions. In *Advances in Automatic Text Summarization*. MIT Press.

Kupiec, J., Pedersen, J. & Chen, F. (1995). A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 68–73). ACM.

Barzilay, R. & Elhadad, M. (1999). Using lexical chains for text summarization. In *Advances in Automatic Text Summarization*. MIT Press.

Boguraev, B. & Kennedy, C. (1999). Salience-based content characterisation of text documents. In *Advances in Automatic Text Summarization*. MIT Press.

Marcu, D. (1999). From discourse structures to text summaries. In: *Advances in Automatic Text Summarization*. MIT Press.

Gong, Y. & Liu, X. (2002). Generic text summarization using relevance measure and latent semantic analysis. In: *Proceedings of ACM SIGIR'02*. ACM.

Erkan, G. & Radev, D. (2004). LexRank: Graph-based centrality as salience in text summarization. In *Journal of Artificial Intelligence Research (JAIR)*.

Hovy, E. (2005). Automated text summarization. In Ruslan Mitkov (ed.), *The Oxford Handbook of Computational Linguistics* (pp. 583–598). Oxford University Press.

Generative summarisation:

McKeown, K. & Radev, D. (1995). Generating summaries of multiple news articles. In: *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 74–82). ACM.

Knight, K. & Marcu, D. (2002). Summarization beyond sentence extraction: A probabilistic approach to sentence compression. In *Artificial Intelligence, 139(1)* (pp. 91–107).

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

Clarke, J. & Lapata, M. (2008). Global inference for sentence compression: An integer linear programming approach. In *Journal of Artificial Intelligence Research, 31* (pp. 273–318).

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

Multilingual/cross-lingual summarisation:

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* (pp. 927–936). ACL.

Boudin, F., Huet S. & Torres-Moreno, J.M. (2010). A graph-based approach to cross-language multi-document summarization. In: *Research journal on Computer science and computer engineering with applications (Polibits), 43* (pp. 113–118).

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

Wan, X., Jia, H., Huang, S., Xiao, J. (2011). Summarizing the differences in multilingual news. In *Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval* (pp. 735–744). ACM.

Steinberger, J., Kabadjov, M., Steinberger, R., Tanev, H., Turchi, M. & Zavarella, V. (2012). Towards language-independent news summarization. In *Proceedings of the Text Analysis Conference 2011*, NIST.

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

Summarisation evaluation:

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

Nenkova, A., Passonneau, R. & McKeown, K. (2007). The pyramid method: incorporating human content selection variation in summarization evaluation. In *ACM Transactions on Speech and Language Processing 4(2).*

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

Owczarzak, K., Conroy, J., Trang Dang, H. & Nenkova, A. (2012). An Assessment of the Accuracy of Automatic Evaluation in Summarization. In *Proceedings of the NAACL Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*. ACL.

Summarisation evaluation in multiple languages:

Giannakopoulos, G., El-Haj, M., Favre, B., Litvak, M., Steinberger, J. & Varma, V. (2012). TAC 2011 Multiling pilot overview. In *Proceedings of the Text Analysis Conference 2011*, NIST.

Turchi, M., Steinberger, J., Kabadjov, M. & Steinberger, R. (2010). Using Parallel Corpora for Multilingual (Multi-Document) Summarisation Evaluation. In: *Multilingual and Multimodal Information Access Evaluation, LNCS 6360* (pp. 52-63), Springer.

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

KEY TERMS & DEFINITIONS

Multilingual summarisation:

It is a summarisation task, in which the languages of the summary and input articles are the same, however, the summariser can process articles in a set of languages.

Cross-lingual summarisation:

It is a summarisation task, in which the languages of the summary and input articles are different.

Aspect-driven summarisation:

It is a summarisation task, in which a summariser is given a list of aspects for each article category, and the summary should include those aspects if possible

Latent semantic analysis:

It is a fully automatic mathematical/statistical technique which is able to extract and represent the meaning of terms on the basis of their contextual usage.

Text Analysis Conferences:

It is a series of evaluation workshops organized by NIST (US National Institute for Standards and Technology) to encourage research in Natural Language Processing and related applications, by providing a large test collection, common evaluation procedures, and a forum for organizations to share their results.

Multilingual summarisation evaluation:

It is a task of evaluating quality of automatically produced summaries in a set of languages.

Parallel corpus

It is a collection of texts placed alongside its translations. Texts and corresponding translations are usually aligned at sentence level.

---

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

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

[iii] http://www.nist.gov/tac/2010/Summarization/Guided-Summ.2010.guidelines.html.

[vi] http://nl.ijs.si/telri/Vanilla/.