JRC's Participation at TAC 2011:
Guided and Multilingual Summarization Tasks

Josef Steinberger,
Mijail Kabadjov, Ralf Steinberger, Hristo Tanev, Marco Turchi, Vanni Zavarella
Joint Research Centre,
European Commission,
Via E. Fermi 2749,
21027 Ispra (VA), Italy
[name].[surname]@jrc.ec.europa.eu

Abstract

The paper describes our participation in the Guided and Multilingual Summarization Tasks at the Text Analysis Conference 2011 (TAC'11). We participated in the Guided task with the system from the previous year which combines aspect identification by an event extraction system and automatically learned lexicons with LSA-based summarizer. This year we included temporal analysis to improve sentence ordering, detection of update information and dealing with the WHEN aspect. We made a first try to compress and paraphrase sentences with our second run. Multilingual summarization is our ultimate goal and thus all components of the system are either fully language independent or can be adapted to other languages relatively easily. The multilingual task provided a possibility to test the system on other languages than English. The sentence-extractive summarizer was ranked among the top systems in readability and non-redundancy. Even if the content of its summaries was not ranked on the top for English in the main Guided task, it reached the top results in the Multilingual task. The generative run suffered from worse readability which affected also the content scores.

1 Introduction

We follow the route towards our main goal – producing multilingual summaries within the Europe Media Monitor (EMM)[1] framework. EMM gathers around 100,000 news articles every day from over 3000 news sources. All news articles are clustered producing topic-homogeneous news clusters for each of the 40+ languages. A summarizer could reduce this big bulk of highly redundant news data. It should be of high quality. However, given the fact we work with so many languages, it has to be enough language-independent to guarantee similar performance across languages. This year TAC included a Multilingual summarization task, a great opportunity to evaluate our system on different languages. This way we could test the language-independent behavior of the summarizer, even on languages we haven't worked with yet.

We participated in the previous TACs. We started in 2008 with the lexical LSA-based approach (Steinberger and Ježek, 2009), which tries to capture and extract the best sentences about the most important concepts in the source articles. In 2009, we included named entities in the summarizer's input representation. In 2010 we participated in the new Guided summarization task. 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. Lately, we started experimenting with sentence (re-)generation (Turchi et al., 2010). The approach lies in between extraction and generation. After extracting the full sentence it compresses it to a sequence of important terms which makes it similar to sentence compression. Finally, the sequence is made more readable by the reconstruction phase based on a language model.

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

This year we participated in the Guided task again. We included analysis of temporal expressions which was used in sentence ordering, dealing with the WHEN aspect and identifying update information. We ran our basic version of the summarizer on all language variants of the Multilingual task to test its language independence. Our previous study (Turchi et al., 2010) in that direction found that even if the summarizer does not use any language-specific properties[2], it selects different sentences for different languages: the summarizer selects on average only 35% of the same sentences for a language pair in a parallel corpus. The fact that, overall, the sentence selection agreement across languages is quite so low indicates that there is a real need for multilingual summarization evaluation.

In the next section we describe our summarization approach, starting with the basic LSA-based method used for the multilingual task followed by the improvements used for the Guided summarization task. At the end of the section we describe the approach of sentence compression/reconstruction. Section 3, resp. 4, discusses results obtained in the Guided, resp. Multilingual, task. Finally, we draw conclusions.

2 Summarization Approach

2.1 Raw LSA-based Approach

Originally proposed by Gong and Liu (2002) and later improved by J. Steinberger and Ježek (2004), this approach 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, we first build matrix **A** where each column represents the weighted term-frequency vector of sentence j in a given set of documents (an initial or update set of documents). The weighting scheme we found to work best is using a binary local weight and an entropy-based global weight (for details see Steinberger and Ježek (2009)).

After that step Singular Value Decomposition (SVD) is applied to the above matrix as $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$, and subsequently matrix $\mathbf{F} = \mathbf{S} \cdot \mathbf{V}^T$ reduced to r dimensions[3] is derived.

Sentence selection starts with measuring the length of sentence vectors in matrix **F**. The length of the vector can be viewed as a measure for importance of that sentence within the top cluster topics. We call it 'co-occurrence sentence score'.

The sentence with the largest score is selected as the first to go to the summary (its corresponding vector in **F** is denoted as $\mathbf{f}_{best}$). After placing it in the summary, the topic/sentence distribution in matrix **F** is changed by subtracting the information contained in that sentence:

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

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 **F**. The process is iteratively repeated until the required summary length is reached.

Our approach to deal with the update problem is to change the weighting scheme in order to give the novel features larger weights. The novelty factor is added to the formula of the weighting scheme (for details see our TAC'09 report (Steinberger et al., 2009a)).

2.2 Entities as Additional Information

Within the EMM's NewsExplorer project[4] R. Steinberger et al. (2009b) developed multilingual tools for geo-tagging (Pouliquen et al., 2006) and entity disambiguation (Pouliquen and Steinberger, 2009). We used both systems to extract information about mentions of the entities in the TAC clusters. The extracted features were used as additional features in co-occurrence calculation generalizing the notion of term but also to capture several aspects (places of events and persons involved in investigations).

[2] Its principle is to use co-occurrence of features in sentences. It ignores words in the stop-word list which is the only language-specific, but easily accessible, resource.

[3] The degree of importance of each 'latent' topic is given by the singular values and the optimal number of latent topics (i.e., dimensions) r can be fine-tuned on training data.

[4] `http://emm.newsexplorer.eu/`

2.3 Event Extraction for Detecting Aspects

NEXUS is an event extraction system which analyzes news articles reporting on violent events, natural or man-made disasters (see Tanev et al. (2008) for 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 Steinberger et al. (2009c).

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 our summarization experiments we ran the event extraction system on each news article from the corpus and we mapped extracted slots to summarization 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 "Victims" were obtained as a union of the event slots: "Dead victims", "Injured", "Arrested", "Displaced", "Kidnapped", "Released hostages" and "People left without homes".

For the other 4 aspects we generated lexicons using Ontopopulis – a system for the automatic learning of semantic classes (see Tanev et al. (2010) for algorithm overview and evaluation). As an input, it accepts a list of words, which belong to a certain semantic class, e.g. "disasters", then it learns additional words, which belong to the same class. Ontopopulis is a multilingual adaptation of a syntactic approach described earlier in Tanev and Magnini (2006). The four aspects covered by our lexicons are: "Damages", "Countermeasures", "Resource", and "Charges". The words and multi-words from these four lexicons were used to trigger the corresponding summary aspects.

We use the identified aspects to boost the co-occurrence-based scores of the sentences that contain them. For each document set we build an aspect-by-sentence matrix which contains Boolean values to store an aspects' presence/absence in sentences[5].

The length of the sentence vector in the aspect matrix works as a booster for the co-occurrence score. After selecting a sentence we lower the influence of the aspects mentioned there. For details see our report from last year (Steinberger et al., 2010).

In the case of update summaries we considered only those aspect mentions that do not occur in the basic documents. For instance, if we find in the basic document set that "200 people were killed", this string is not considered as a mention of the AFFECTED aspect if found in the update document set. However, if there is more specific information like "212 people were killed" the aspect in the sentence is turned on.

2.4 Temporal Analysis

Temporal information is variously encoded by different grammar features in news text, including tense and aspect markers, temporal clausal conjunctions, adverbial and prepositional phrases. We focus our temporal analysis on a specific subset of linguistic constructions, the so-called temporal expressions (timex), whose extent is approximately defined as in Ferro et al. (2005). They are characterized by linear insulation and the presence of one from a finite set of lexical triggers (e.g. 'Monday', 'November','yesterday', 'hours', '2010' etc.), which make them easy to model by finite state grammars.

The time expressions we model range across several dimensions: they include the numerical vs. non-numerical format ('03/18/2010', 'on the fifth of December 2009'), fully specified vs. underspecified ('on the fifth of December 2009', 'in March 2002'), absolute vs. relative vs. deictic ('in March 2002', 'in March', 'last month'), simple vs. compound ('a year before last Monday'), discrete vs. fuzzy ('three days ago', 'in a few months'). Regarding the classification of temporal entities, we distinguish dates, periods, durations and time sets.

[5] Only aspects relevant for the topic category are taken.

The Temporal Information extraction module we deployed consists of two stages: Recognition and Normalization. In the Recognition phase, timexes are detected and segmented in text and a more abstract representation of them is filled for further processing. It deploys a number of text processing modules from the CORLEONE tool set (Piskorski, 2008), including tokenization, morphological and temporal lexicon lookup. Local parsing of timexes is performed by a manually designed cascade of partially language-independent finite-state grammar rules using the EXPRESS pattern matching engine (Piskorski, 2007), resulting in an intermediate feature structure-like representation to be used by the normalization module. This consists firstly of "anchor selection", that is determining and maintaining a reference time for relative timex resolution, which starts by using the article date from the TAC data and updates it along the resolution process according to some heuristics. Then, the reference time is used to resolve relative timexes, computing their actual values via calendar arithmetic. Finally, their representation is normalized according to a machine-readable standard (we use a variant of Timex2 (Ferro et al., 2005)).

We used the normalized temporal expressions for three different tasks in the summarization process. First, it was used for the detection of the WHEN aspect. The most frequent normalized expression was taken as the time of the accident/attack. Secondly, we ordered sentences according to the temporal expressions. Each sentence was attached to a date it contained. If a sentence contained more temporal expressions the first one was selected to simplify the problem. If there was no timex in the sentence it looked back to find the last one given the hypothesis that each topic is introduced by a sentence which refers to a date/time, followed by details of the event. Usually, the first sentences in news articles mention a date/time, however if they do not, we take the date of the article.

The last usage of timex information is in identifying the update sentences. If a sentence mentions a date which is more recent than the date of the most recent article in the initial set of documents then it probably reports on an event that happened in a later time period than the initial set. The approach is the following. We use as reference the publication date of the most recent article in the initial set of documents. Then for each sentence of the update set, we normalize all the timexes (if any), then we convert both into an interval representation so we can compute pairwise interval relations as defined by Allen's Interval Algebra (Allen, 1983). Update sentences are the ones in which at least one of the temporal intervals is in an "after", "overlapped_by" or "finishes" relation with the reference one. Notice that we do not keep information about events referenced in text, so that anaphoric references to events are not resolved. A limitation following from this is that event-event temporal relation markers (e.g. 'after the attack on Bagdad') cannot be used in the update detection task, therefore potentially reducing the recall of the method.

2.5 Sentence Compression and Reconstruction

Empirical evidence shows that human summaries contain on average more and shorter sentences than the system summaries. 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. We have already tried to investigate language-independent possibilities in that direction (Turchi et al., 2010). Our initial experimental results showed that our approach is feasible, since it produced summaries, which when evaluated against the TAC 2009 data yield ROUGE scores comparable to the average of the participating systems. However, it achieved lower scores compared to our sentence-extractive summarizer.

The approach starts with identifying the most salient terms in each selected sentence. For each term we compute the term salience score from the LSA[6] and language model probabilities up to 4-grams. 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 normalizing scores of each feature and combining them, each term ended up with a score reflect-

[6] The magnitude of its corresponding vector in the matrix resulting from the dot product of the matrix of left singular vectors with the diagonal matrix of singular values. More formally, let $T = U \cdot \Sigma$ and then for each term i, the salience score is given by $|\vec{T_i}|$.

Run ID	Overall responsiveness	Linguistic quality	Pyramid score	No. of repetitions
25 (the best run in Overall resp.)	**3.159** (1)	3.341 (6)	0.44 (10)	1.409 (17/25)
22 (the best run in Pyramid score)	3.136 (2)	3.432 (5)	**0.477** (1)	1.045 (7/25)
37 (sentence extraction)	2.977 (12)	**3.455** (4)	0.412 (23)	**0.864** (2/25))
6 (+ compression/paraphrasing)	2.341 (43)	2.318 (42)	0.311 (42)	0.568 (-/25)
2 (baseline - MEAD)	2.841 (27)	2.818 (30)	0.362 (32)	1.432 (-/25)
1 (baseline - LEAD)	2.5 (37)	3.205 (7)	0.304 (45)	0.455 (-/25)

Table 1: TAC'11 results of the Guided summarization task - initial summaries.

Run ID	Overall responsiveness	Linguistic quality	Pyramid score	No. of repetitions
35 (the best run in Overall resp.)	**2.591** (1)	2.818 (24)	0.342 (4)	0.818 (19/25)
9 (the best run in Pyramid score)	2.523 (5)	2.659 (34)	**0.353** (1)	0.409 (3/25)
37 (sentence extraction)	2.205 (31)	3.25 (6)	0.291 (21)	**0.25** (1/25)
6 (+ compression/paraphrasing)	1.864 (45)	2.159 (44)	0.176 (44)	0.295 (-/25)
2 (baseline - MEAD)	2.114 (35)	2.841 (22)	0.284 (24)	0.568 (10/25)
1 (baseline - LEAD)	2.091 (37)	**3.455** (1)	0.237 (36)	0.364 (-/25)

Table 2: TAC'11 results of the Guided summarization task - update summaries.

ing 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 (for details see Turchi et al. (2010)), adding the most probable content (mainly stopwords) to connect the sentence fragments. The term selection gives compression capabilities and the reconstruction adds paraphrasing capabilities.

3 Guided Summarization Task

The task was the same as last year: to write a 100-word summary for a set of 10 newswire articles for a given topic, where the topic falls into pre-defined categories. Participants were given a list of important aspects for each category, and the summary should cover all those aspects if possible. The summaries could also contain other information relevant to the topic. There was also the update part of the task: write a 100-word update summary of a subsequent 10 newswire articles for the topic, under the assumption that the user has already read the earlier articles.

The first run submitted for the guided task is similar to the one submitted last year, however, we made several changes. Analyzing our TAC'10 results we found out that giving more weight to sentences that contained certain aspects led to select better sentences but in the case of other aspects the influence was negative. In the case of the negative effect, we found out that either the aspect definition is too wide (e.g., the WHEN aspect – giving advantage to sentences with whatever date), and should be focused on the most frequent aspect mention, or treating the aspect the way we did is not helpful at all. Thus, we trained on last year's data the aspect-based part of the summarizer – for each aspect we learned if it is better to take only the most frequent aspect mention or to take all aspect mentions or to turn the aspect off. Another change was to treat the WHEN aspect by the proper temporal analysis. The analysis was used also for sentence ordering and identification of update information. The resulting system was submitted as run1.

For our second run we applied our compression/paraphrasing method (see section 2.5) on the output of run1. Because it resulted in a shorter summary, additional sentences were added (and compressed) to reach the summary limit and not to get the recall handicap.

Tables 1 and 2 contain the overall TAC results for initial and update summaries. We report the results and corresponding ranks (in brackets) within all the 50 systems of the two best TAC systems, our two submissions, and the two baselines.

In the case of initial summaries the sentence-extractive run (37) performed well in linguistic quality. It was ranked 4th, surprisingly higher than the baseline, which selects a continuous text – the beginning of the most recent article. Content was a little bit above average (23rd). Altogether it resulted in 12th rank in overall responsiveness. Number of repetitions showed low redundancy – 2nd best among the top performing systems[7].

In the case of update summaries the linguistic quality was high again (6th). The Pyramid score indicates above average content (21st). Surprisingly, the overall responsiveness does not fall between the rank of ling. quality and the Pyramid score, ranking our extractive approach 31st.

The second run (6) performed worse in linguistic quality as expected (42nd for initial summaries, 44th for update summaries), but also significantly worse in the content-based Pyramid method (43rd / 44th) showing that still the sentence-extractive approach performs better than the generative one.

Compared to last year the system performed significantly better in capturing the WHEN aspect in update summaries and slightly better in linguistic quality indicating the positive contribution of the temporal analysis.

4 Multilingual Task

The Multilingual task aims to evaluate the application of (partially or fully) language-independent summarization algorithms on a variety of languages. We used our basic version of the summarizer (Section 2.1) to create summaries for all the seven languages of the task. The only resource dependent on the language was a list of stopwords. We did not use our entity detection, event extraction and temporal analysis tools because we haven't developed them yet for all the languages of the task. However, we are working on their adaptation for other languages[8].

The aim was to generate 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. Important difference from the main TAC summarization task was that the limit of summary length was set to 250 words. For our LSA-based system it meant to raise the dimensionality reduction to include more latent dimensions (= more topics). Because we did not perform temporal analysis for this task we ordered sentences in the summary based on the date of the document they came from, sentences from the same document followed their order from the full text. Even if they were sometimes out of context, when extracted, the adjacent sentences at least dealt with the same event. For readability evaluation, sentence order is important for event-based stories.

In the following tables we compare our system (LSA-based summarizer), baseline (Centroid Baseline – the start of the centroid article), topline (GA Topline – summary based on genetic algorithm using model summary information), the best summarizer for each language (if our system was the best we report the result of the second best one). In total 10 systems participated. We show scores and ranks per language[9] and averages for all languages[10].

Human annotators scored each summary on the 5-to-1 scale (5 = the best, 1 = the worst) – human grades (table 3). The score corresponds to the overall responsiveness of the main TAC task – equal weight of content and readability. To avoid advantage of shorter summaries, which humans could score higher, their grade was scaled down – length-aware human grades (table 4). We report also figures of 2 automatic measures: the AutoSummENG metric (Merged Model Graphs variation (Giannakopoulos and Karkaletsis, 2010)) and the ROUGE-2 score[11] (Lin, 2004) in tables 5 and 6.

In the case of raw human evaluation our system was ranked at the top for 5 languages – Czech (our system was baseline there), English, French, Hebrew and Greek. For Arabic it was lower than baseline and for Hindi three other systems performed

[7] We compare only the top half of the systems based on Pyramid score because taking in the low performing systems would distort it. An empty summary is perfect from the point of view of number of repetitions.

[8] So far our NER works for 20 languages (ar, bg, es, en, et, da, de, fa, fr, it, nl, no, pl, pt, ro, ru, sl, sv, ew, tr), event extraction for 8 languages (ar, de, en, es, fr, pt, ru, tu) and temporal analysis for 4 languages (en, es, fr, it).

[9] We were coordinators of the Czech language evaluation, thus our system serves only as baseline for Czech.

[10] Only 7 systems participated in all languages.

[11] We found similar rankings in the case of ROUGE-SU4.

Run	Arabic	Czech	English	French	Hebrew	Hindi	Greek	Average
The best run	ID1	ID1	ID2	ID1	ID1&ID2	ID5	ID2	ID1
(excl. ours and baselines)	**3.77**	3.00	3.53	2.30	3.29	**2.73**	3.33	3.01
	(1/9)	(3/7)	(2/10)	(2/9)	(2/7)	(1/9)	(2/7)	(2/7)
ID3 (our system)	3.43	*3.40*	**3.57**	**3.23**	**3.87**	2.47	**3.63**	**3.37**
	(4/9)	*(1/7)*	(1/10)	(1/9)	(1/7)	(4/9)	(1/7)	(1/7)
ID9 (Centroid Baseline)	3.73	**3.30**	2.27	2.03	3.16	1.80	3.13	2.78
	(2/9)	(2/7)	(8/10)	(7/9)	(4/7)	(8/9)	(4/9)	(4/7)
ID10 (GA Topline)	3.20	2.68	3.20	2.10	3.03	1.83	3.30	2.76
	(6/9)	(5/7)	(3/10)	(5/9)	(6/7)	(7/9)	(3/9)	(5/7)

Table 3: Average **human grades** (5 = the best, 1 = the worst) and ranks of the systems (rank/number of participating systems).

Run	Arabic	Czech	English	French	Hebrew	Hindi	Greek	Average
The best run	ID1	ID1	ID2	ID1	ID1&ID2	ID5	ID2	ID1
(excl. ours and baselines)	**3.77**	3.00	**3.53**	2.28	3.29	**2.60**	**3.33**	2.99
	(1/9)	(3/7)	(1/10)	(2/9)	(2/7)	(1/9)	(1/7)	(2/7)
ID3 (our system)	3.10	*3.15*	3.31	**2.93**	**3.56**	2.20	3.25	**3.07**
	(7/9)	*(2/7)*	(2/10)	(1/9)	(1/7)	(4/9)	(3/7)	(1/7)
ID9 (Centroid Baseline)	3.73	**3.30**	2.27	2.03	3.16	1.80	3.13	2.78
	(2/9)	(1/7)	(7/10)	(6/9)	(4/7)	(6/9)	(4/9)	(4/7)
ID10 (GA Topline)	3.20	2.68	3.20	2.10	2.85	1.75	3.30	2.73
	(5/9)	(5/7)	(3/10)	(4/9)	(6/7)	(7/9)	(2/9)	(5/7)

Table 4: Average **human grades scaled down for shorted summaries** (5 = the best, 1 = the worst) and ranks of the systems (rank/number of participating systems).

Run	Arabic	Czech	English	French	Hebrew	Hindi	Greek	Average
The best run	ID4	ID2	ID2	ID2	ID2	ID2	ID2	ID2
(excl. ours and baselines)	0.383	**0.373**	0.386	0.414	0.327	**0.286**	**0.375**	0.361
	(2/8)	(2/6)	(2/9)	(2/8)	(2/6)	(1/8)	(1/6)	(2/6)
ID3 (our system)	**0.483**	*0.430*	**0.426**	**0.466**	**0.368**	0.275	0.372	**0.403**
	(1/8)	*(1/6)*	(1/9)	(1/8)	(1/6)	(2/8)	(2/6)	(1/6)
ID9 (Centroid Baseline)	0.282	0.312	0.304	0.336	0.272	0.207	0.291	0.286
	(7/8)	(6/6)	(9/9)	(8/8)	(6/6)	(6/8)	(5/6)	(6/6)
ID10 (GA Topline)	0.666	0.689	0.548	0.595	0.537	0.361	0.524	0.560

Table 5: Average scores of the **AutoSummENG** metric and ranks of the systems (rank/number of participating systems). Ranks do not take into account the Topline baseline which uses model summaries as its input, and thus, it is the most similar to them.

Run	Arabic	Czech	English	French	Hebrew	Hindi	Greek	Average
The best run	ID8	ID2	ID2	ID2	ID2	ID5	ID2	ID2
(excl. ours and baselines)	0.147	**0.190**	0.171	0.197	0.095	0.034	**0.149**	0.136
	(2/8)	(2/6)	(2/9)	(2/8)	(3/6)	(2/8)	(1/6)	(2/6)
ID3 (our system)	**0.158**	*0.199*	**0.173**	**0.202**	**0.129**	**0.058**	0.101	**0.146**
	(1/8)	*(1/6)*	(1/9)	(1/8)	(1/6)	(1/8)	(3/6)	(1/6)
ID9 (Centroid Baseline)	0.126	0.140	0.110	0.130	0.102	0.000	0.073	0.097
	(6/8)	(5/6)	(7/9)	(6/8)	(2/6)	(6/8)	(4/6)	(5/6)
ID10 (GA Topline)	0.234	0.351	0.252	0.286	0.222	0.000	0.145	0.213

Table 6: Average scores of the **ROUGE-2** metric and ranks of the systems (rank/number of participating systems). Ranks do not take into account the Topline baseline which uses model summaries as its input, and thus, it is the most similar to them.

better. Looking at the average across languages, our system received a promising value, 3.37, in front of the system with ID1 (3.01). The baseline was better than two participating systems and worse than three. Even if Topline uses information from model summaries it was ranked behind the baseline[12].

Because our summaries were several times shorter than the limit (240 words), their scores went down in the case of length-adjusted human grades. They were shorter only when the summarizer would select a long sentence that would result in crossing the 250-word limit. However, even the average over all languages of the lowered scores was ranked at the top (3.07), although there were insignificant differences compared to two following systems (ID1 - 2.99 an ID2 - 2.96). Our system performed the best in two languages: French and Hebrew.

Looking at the results of the AutoSummENG metric the distance between our system and the 2nd best system was even larger. Our system dominated in 5 languages and twice was ranked as second. The average for all languages was 0.403. The second best system (ID2) received 0.361. All the systems were far away from the topline (0.560), however, it has only indicative role in our figures because it uses information from model summaries.

ROUGE-2 found similar results as the AutoSummENG metric did. Our system was ranked 1st in 6 languages, only in Greek it was ranked 3rd. Low numbers in Greek and Hindi indicate that ROUGE has to be set up differently or it is not suitable for evaluation of these languages. Given the fact that both the automatic metrics evaluate summary content (recall-based) it seems that content of our summaries was good enough even if they were sometimes shorter than the limit (240 words). It indicates better correlation with NOT length-adjusted human grades.

5 Conclusion

We participated in two tasks of the summarization track. In the main guided summarization task our sentence-extractive run performed well in readability and above average in content. The results were similar to our participation last year. There were several indications of positive effects of the new temporal analysis we included. The generative run suffered from worse readability, but also in the case of content the changes in sentences led to losing important SCUs according to the lower Pyramid scores. In the multilingual task our basic system performed really well. It was ranked the best even in languages we are not familiar with at all (e.g. Hebrew). To conclude, we found the performance of our LSA-based summarizer sufficiently good in English, and the fact that it does not use any language-specific tools/resources and thus can be run for other languages led to the top scores in the multilingual task.

References

J.F. Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26:832–843.

L. Ferro, L. Gerber, I. Mani, B. Sundheim, and G. Wil-

[12] Given the fact that the multilingual pilot consisted only of 10 topics there were not many statistically significant differences between the systems.

son. 2005. *TIDES 2005 Standard for the Annotation of Temporal Expressions*.

G. Giannakopoulos and V. Karkaletsis. 2010. Summarization system evaluation variations based on n-gram graphs. In *Proceedings of the Text Analysis Conference (TAC)*.

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

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

J. Piskorski. 2007. Express extraction pattern recognition engine and specification suite. In *Proceedings of the International Workshop Finite-State Methods and Natural language Processing*.

J. Piskorski. 2008. Corleone: Core linguistic entity online extraction. Technical Report EUR23393, Joint Reserach Centre.

B. Pouliquen and R. Steinberger. 2009. Automatic construction of multilingual name dictionaries. In Cyril Goutte, Nicola Cancedda, Marc Dymetman, and George Foster, editors, *Learning Machine Translation*. MIT Press, NIPS series.

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

J. Steinberger and K. Ježek. 2004. Text summarization and singular value decomposition. In *Proceedings of the 3rd ADVIS conference*, Izmir, Turkey.

J. Steinberger and K. Ježek. 2009. Update summarization based on novel topic distribution. In *Proceedings of the 9th ACM Symposium on Document Engineering, Munich, Germany*.

J. Steinberger, M. Kabadjov, B. Pouliquen, R. Steinberger, and M. Poesio. 2009a. WB-JRC-UTs participation in tac 2009: Update summarization and aesop tasks. In *Proceedings of the Text Analysis Conference (TAC)*.

R. Steinberger, B. Pouliquen, and C. Ignat. 2009b. Using language-independent rules to achieve high multilinguality in text mining. In François Fogelman-Soulié, Domenico Perrotta, Jakub Piskorski, and Ralf Steinberger, editors, *Mining Massive Data Sets for Security*. IOS-Press, Amsterdam, Holland.

R. Steinberger, B. Pouliquen, and E. Van der Goot. 2009c. An introduction to the europe media monitor family of applications. In *Information Access in a Multilingual World Proceedings of the SIGIR*.

J. Steinberger, H. Tanev, M. Kabadjov, and R. Steinberger. 2010. Jrc's participation in the guided summarization task at tac 2010. In *Proceedings of the Text Analysis Conference (TAC)*.

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

H. Tanev, J. Piskorski, and M. Atkinson. 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)*.

H. Tanev, V. Zavarella, J. Linge, M. Kabadjov, J. Piskorski, M. Atkinson, and R. Steinberger. 2010. Exploiting machine learning techniques to build an event extraction system for portuguese and spanish. *Journal Linguamatica: Revista para o Processamento Automatico das Linguas Ibericas*.

M. Turchi, J. Steinberger, M. Kabadjov, R. Steinberger, and N. Cristianini. 2010. Wrapping up a summary: from representation to generation. In *Proceedings of CLEF*.