WB-JRC-UT's Participation in TAC 2009: Update Summarization and AESOP Tasks*

Josef Steinberger[1,2], Mijail Kabadjov[2], Bruno Pouliquen[2], Ralf Steinberger[2], and Massimo Poesio[3,4]

[1] University of West Bohemia, Univerzitni 8, Pilsen 306 14, Czech Republic
jstein@kiv.zcu.cz
[2] Joint Research Centre, European Commission, Via E. Fermi 2749, Ispra (VA), Italy
firstname.lastname@jrc.ec.europa.eu
[3] University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom
poesio@essex.ac.uk
[4] Universitá di Trento, Povo, TN 38100, Italy

Abstract. In this paper we describe our participation in the Summarization track at the 2009 Text Analysis Conference (TAC). The Summarization track was composed of two tasks: Update Summarization and Automatically Evaluating Summaries of Peers (AESOP). We submitted two runs for the former and four for the latter. In the following sections we describe our runs and discuss the results attained.

1 Introduction

News collators such as Google News, Yahoo News or the Europe Media Monitor (EMM)[5] gather hundreds of thousands of news articles every day from multiple sources. Every ten minutes, EMM's applications NewsBrief and the Medical Information System MedISys cluster the news collected during the last few hours. This may result in tens, if not hundreds of articles about the same event or subject. Multi-document summarisation is a potentially promising way to reduce this big bulk of highly redundant news data and obtain one easily consumable text summarizing the most important content.

News stories on events typically live for a few days only, but in some cases–eg., electoral campaigns–stories may live for weeks or even months. In either case, the news content usually changes significantly from the first event reports to the articles published later on. In the case of events such as disasters or accidents, victim counts typically change and eyewitness reports may provide additional details. In the case of election rallies, new election topics may come into focus and popularity measures may change. Update summaries have the objective of providing details on the changes since earlier reports, assuming that the readers are aware of the initial event.

* We would like to thank the EMM team for making available to us their robust tools for multilingual text mining.

[5] http://emm.jrc.it/overview.html

The TAC 2009 competition provides the infrastructure to compare various methods to produce both initial and update summaries, as well as to automatically evaluate various multi-document summarisation methods. In this paper, we present our approach to multi-document summarisation (section 2) and to the automatic evaluation of summaries (section 3), and we discuss the results achieved in TAC 2009. In Section 4, we summarise the contents of this paper.

2 Update Summarization

The TAC 2009 Update summarization task required systems to produce a short (100 words) summary of a set of newswire articles (an update summary), under the assumption that the user has already read a given set of earlier articles (used for the creation of an initial summary). Our update summarizer was obtained by merging ideas developed in two separate strands of earlier work. We first briefly describe these, then explain how they were combined before illustrating the results.

2.1 Two Earlier Strands of Work

LSA-Based Summarization. Originally proposed by Gong and Liu (2002) and later improved by 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 $A = [A_1, A_2, \ldots, A_n]$, where each column $A_j = [a_{1j}, a_{2j}, \ldots, a_{nj}]^T$ represents the weighted term-frequency vector of sentence j in a given set of documents. Each element in this vector is defined as:

$$a_{ij} = L(i,j) \cdot G(i), \tag{1}$$

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 whole set of documents. The weighting scheme we found to work best is using a binary local weight and an entropy-based global weight:

$$\begin{array}{ll} L(i,j) = 1 & \text{if term } i \text{ appears at least once in sentence } j; \\ L(i,j) = 0 & \text{otherwise.} \end{array} \tag{2}$$

$$G(i) = 1 - \sum_j \frac{p_{ij}\log(p_{ij})}{log(n)}, p_{ij} = \frac{f_{ij}}{g_i}, \tag{3}$$

where f_{ij} 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. If there are m terms and n sentences in the set, an $m \times n$ matrix A will be obtained.

After that step Singular Value Decomposition (SVD) is applied to the above matrix as $A = U \Sigma V^T$, and subsequently sentences are extracted by the iterative approach based on sentence vectors' lengths in matrix $\Sigma \cdot V^T$ reduced to r dimensions (for details see (Steinberger and Ježek, 2009)).[6]

The aim of update summarization is to produce a summary from the set of recent documents under the assumption that the reader has already read a set of older documents concerned with the same topic. Our approach is to change the weighting scheme in order to give the novel features larger weights. *Novelty*, N, is added to formula (1):

$$a_{ij} = L(i,j) \cdot G(i) \cdot N(i), \tag{4}$$

where L and G are the local and global weights in the set of recent documents. We applied the following formula to compute the novelty weights:

$$N(i) = \begin{cases} nov_{max}; \quad if \quad g_i^{(old)} = 0 \\ min \left[nov_{max}, 1 + \log \left(\frac{g_i^{(old)} + g_i^{(new)}}{g_i^{(old)}} \right) \right]; \, otherwise \end{cases} \tag{5}$$

where $g_i^{(old)}$ is the frequency of term i in the set of older documents, $g_i^{(new)}$ is the frequency of term i in the set of recent documents, and nov_{max} is the maximal weight increase for novel terms.

Steinberger et al. (2007) enhanced the above approach for single-document summarisation by augmenting the source matrix A with extra dimensions generalising the notion of 'term' to be also mentions of entities from the real world. The extra entity-by-sentence dimensions were obtained using the anaphora resolution system GuiTAR[7].

Multilingual Entity Disambiguation for Cross-lingual News Cluster Linking. Working on a different problem –cross-lingual linking of news clusters within the EMM's NewsExplorer project[8]– R. Steinberger et al. (2009) also developed a language-independent representation combining various sources of knowledge. As in the work of J. Steinberger et al. (2007), the EMM representation consisted of both keywords and terms. In this case, the terms included geographical locations, person and organisation entities, and EUROVOC indexing. For this work, R. Steinberger et al. (2009) developed multilingual tools for geo-tagging Pouliquen et al. (2006) and entity disambiguation (Pouliquen and Steinberger, 2009).

[6] 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.

[7] `http://guitar-essex.sourceforge.net/`

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

2.2 LSA-Based Summarization meets Entity Disambiguation

For the TAC 09 update summarizer, we used the NewsExplorer multilingual tools for geo-tagging and entity disambiguation developed by R. Steinberger *et al.* and used them to augment the source entity-by-sentence matrix A used in the LSA-based summariser proposed by J. Steinberger *et al.*. In addition, we augmented the matrix with terms grounded to the Medical Subject Headings (MeSH) taxonomy, taking advantage of the fact that tools for this task were available to us as part of the European Media Motor (EMM) project at the JRC. The main idea behind this is to capture more complex semantic relationships such as hypernymy and synonymy.

2.3 Experimental Results

52 automatically created sets of summaries were submitted by 27 participating groups and compared against three baselines. Baseline 1 (*run1*) returns all the leading sentences (up to 100 words) in the most recent document. This baseline provides a lower bound on what can be achieved with a simple fully automatic extractive summarizer. Baseline 2 (*run2*) returns a copy of one of the model summaries for the docset, but with the sentences randomly ordered. It provides a way of testing the effect of poor linguistic quality on the overall responsiveness of an otherwise good abstractive summary. Baseline 3 (*run3*) returns a summary consisting of sentences that have been manually selected from the docset. It provides an approximate upper bound on what can be achieved with a purely extractive summarizer.

The NIST assessors assigned a content score using Columbia University's Pyramid method (Nenkova and Passonneau, 2004), a readability score and an overall responsiveness score (combining both of the previous ones) to each of the automatic and human summaries. The score is an integer between 1 (very poor) and 10 (very good). Standard automatic scores ROUGE-2, ROUGE-SU4, and BE were calculated as well.

We submitted two runs. For our first priority run (*run19*) we used all types of features: lexical (we used unigrams and bigrams), entity (the output of the entity disambiguation systems) and MeSH terms. In our second run (*run11*) we used all types of features except the MeSH-based ones, to evaluate the contribution of taxonomic information.

Next, we discuss first the results of initial summaries (Table 1) followed by the results of update summaries (Table 2). The top two rows show the scores of the two upper bound baselines (*run2* and *run3*) and the last row corresponds to the lower bound baseline (*run1*). Below the upper bounds we show the results of the best systems evaluated by overall responsiveness (*run40* scored highest in the case of initial summaries and *run24* in the case of update summaries). Results of our runs can be found below (*run19* and *run11*).

Our two runs received very good scores for initial summaries. Our *run19* was the best run overall in linguistic quality, and *run11* was second. *Run19* also received the second highest score in overall responsiveness. (*run11* was 7th in this case.) *run11* did better according to the Pyramid score, 11th; *run19* was 19th. In the case of update summaries, our *run19* was evaluated again as better than *run11* - 9th in overall resp./14th in ling. quality/8th in Pyramids, compared to 13th/14th/13th. The fact that *run19* scored higher than *run11* in all human-based scores except for Pyramids with initial summaries suggests that using taxonomic information has a positive impact on summary quality, although the differences between the runs are not statistically significant. Automatic measures did not seem to correlate well with human-based measures this year.

Run ID.	Overall responsiveness	linguistic quality	Pyramid score	ROUGE-2	ROUGE-SU4	BE
2	6.364	5.477	0.646	0.331	0.344	0.248
3	6.341	7.477	0.358	0.106	0.138	0.053
40	**5.159**	5.636	**0.383**	**0.121**	**0.151**	**0.064**
24	4.955	5.682	0.316	0.098	0.133	0.056
19	4.955	**5.932**	0.277	0.094	0.129	0.052
11	4.795	5.773	0.314	0.096	0.130	0.054
1	3.636	6.705	0.175	0.063	0.099	0.029

Table 1. TAC'09 results of the Update summarization task - initial summaries.

Run ID	Overall responsiveness	linguistic quality	Pyramid score	ROUGE-2	ROUGE-SU4	BE
2	6.182	5.886	0.690	0.319	0.337	0.250
3	6.114	7.250	0.329	0.097	0.136	0.057
40	4.568	5.500	0.290	**0.104**	**0.140**	0.062
24	**5.023**	**5.886**	**0.296**	0.096	0.135	**0.064**
19	4.318	5.182	0.266	0.077	0.116	0.045
11	4.227	5.182	0.247	0.083	0.121	0.047
1	4.318	6.455	0.160	0.051	0.091	0.024

Table 2. TAC'09 results of the Update summarization task - update summaries.

3 Automatically Evaluating Summaries of Peers

In this section we describe the four evaluation metrics we submitted to the task of Automatically Evaluating Summaries Of Peers (AESOP).

3.1 Information Content for Summary Evaluation

We propose to measure the amount of content shared between a pair of texts (e.g., summaries) on the basis of the average semantic similarity between the set of concepts within the first (model) text and the set of concepts within the second text. More formally,

$$avg\_sim(C_m, C_s) = \sum_{c_m \in C_m, c_s \in C_s} \frac{\max\limits_{c_m, c_s} [sim(c_m, c_s)]}{|C_m|} \qquad (6)$$

where C_m is the set of concepts contained in the model summary, C_s is the set of concepts within the system summary, $|C_m|$ denotes the size of C_m[9] and $sim(c_m, c_s)$ is Resnik's semantic similarity measure using a taxonomy (see (Resnik, 1995) for more details).

Furthermore, the above information content-based metric can be easily combined with surface level features such as unigram and bigram recall using a weighted linear combination as follows:

$$score = \alpha \cdot unigrams(M, Sys) + \beta \cdot bigrams(M, Sys) + \gamma \cdot avg\_sim(C_m, C_s) \quad (7)$$

where $unigrams(M, Sys)$ and $bigrams(M, Sys)$ represent the recall of unigrams and bigrams, respectively, of the system summary (Sys) with respect to the model summary (M). In order to ensure an overall score within $[0, 1]$ we set $\alpha + \beta + \gamma = 1$. We estimated the optimal values for these weights on the training data[10] and the best combination obtained was $\alpha = 0.4$, $\beta = 0.4$, $\gamma = 0.2$.

Intuitively, such a hybrid evaluation metric captures on one hand higher level lexical relationships such as hypernymy and synonymy which on the other hand are complemented by direct lexical co-occurence.

3.2 An Alternative Use of the IS-A Taxonomy and Addition of Entities

We experimented with two other evaluation metrics. The main difference with respect to the above information content-driven metric is a different operationalisation of the use of the taxonomy. Instead of defining a similarity measure between concepts, we first expand all taxonomy concepts found in each summary (model and summary) with all their IS-A ancestors (e.g., ... *pneumonia→lung deseases→respiratory tract deseases...*). Then we cast the problem as co-occurence of concepts.

In order to expand further the information we capture we also used a named entity disambiguator (Pouliquen and Steinberger, 2009) and a geo tagger (Pouliquen

[9] By dividing by the number of concepts contained in the model summary, we are aiming at a recall-like metric (as opposed to precision-like, i.e., by dividing by $|C_s|$).

[10] We used the TAC 2008 data for training.

et al., 2006) to identify and disambiguate persons, organisations and geographical locations. Thus we put together all these sources of information by linear combination as follows:

$$score = \alpha \cdot uni(M, Sys) + \beta \cdot bi(M, Sys) + \gamma \cdot cpts(M, Sys) + \delta \cdot ents(M, Sys) \quad (8)$$

where $uni(M, Sys)$ is the recall of unigrams, $bi(M, Sys)$ is that of bigrams, $cpts(M, Sys)$ is the recall of taxonomy concepts and $ents(M, Sys)$ of entities. Again, we set $\alpha + \beta + \gamma + \delta = 1$.

Note that there are two possibilities here for combining recall: one micro-averaged as in eq. 8 and one macro-averaged (i.e., by first summing up all the numerators and denominators separately and then dividing once at the end – both sums can still be weighted as in eq. 8). In fact, on our training data we found that macro-averaging worked better.

3.3 Evaluation

In total, 35 different metrics were submitted by 12 participants for the AESOP task. Two baseline metrics were also used, ROUGE-SU4 and Basic Elements (BE), for a total of 37 runs. All these metrics were assessed in two ways, by using three statistical correlation measures (Pearsons r, Spearmans ρ and Kendalls τ) and by contingency tables specifically highlighting discriminative power. Both of these were measured with respect to two human-produced scores: pyramid score (content) and overall responsiveness (content and readability). In addition, there were two other evaluation dimensions: one including system and model summaries (i.e., all peers) and another one including only system summaries. Finally, initial summaries and update summaries were evaluated separately along all of the above criteria adding up for a total of 32 different evaluation dimensions.

We submitted 4 different metrics, two of which were described in the previous section. We discuss first the evaluation based on statistical correlations, then the one using discriminative power.

Evaluation based on statistical correlations. In Table 3 we present Pearson's correlations[11] of the AESOP metrics with the Pyramid and Overall Responsiveness scores for TAC 2009 initial and update summaries including the model summaries (i.e., all peers).

The top three rows of Table 3 show the metrics that scored highest on all four distinct values. The bottom two rows show the the two baseline metrics. The top three and the bottom two rows are included for reference. The rows in between show the results for our metrics: IC is the information content-based metric, whereas IC+n-grams is the linear combination of that metric with n-gram recall.

[11] All with $p < 0.01$.

Run No.	Initial Summaries		Update Summaries	
	Pyramid	Resp.	Pyramid	Resp.
run11	0.982	**0.968**	0.976	**0.963**
run17	**0.983**	0.963	0.973	0.957
run24	0.978	0.938	**0.978**	0.929
run3, IC	0.771	0.758	*0.701*	0.686
run4, $uni + bi + cpts + ents_1$	0.802	0.701	0.752	0.622
run13, $uni + bi + cpts + ents_2$	0.781	0.674	0.756	0.614
run27, IC+n-grams	0.826	0.74	0.799	0.686
ROUGE-SU4	0.734	0.617	0.726	0.564
BE	0.586	0.456	0.629	0.447

Table 3. Pearson's correlations including models.

Our metrics surpassed the two baselines in all cases except for IC's correlation with the pyramid score for the update summaries (see number in italics). Our best metric for initial summaries, pyramid score was *IC+n-grams* and was ranked 14th out of 37. Our best metric for update summaries, pyramid score was *IC+n-grams* and was ranked 13th out of 37. Our best metric for initial summaries, overall responsiveness score was *IC* and was ranked 13th out of 37. Our best metrics for update summaries, overall responsiveness score were *IC* and *IC+n-grams* and were ranked 14th out of 37.

Run No.	Initial Summaries		Update Summaries	
	Pyramid	Resp.	Pyramid	Resp.
11	0.954	0.829	**0.97**	0.796
24	0.963	0.851	0.957	**0.833**
26	**0.978**	**0.872**	**0.97**	0.814
3	*0.683*	*0.709*	*0.639*	*0.638*
4	0.967	0.851	0.946	0.801
13	0.952	0.809	0.962	0.768
27	0.951	0.854	*0.934*	0.798
1	0.921	0.767	0.94	0.729
2	0.857	0.692	0.924	0.694

Table 4. Pearson's correlations excluding models.

In Table 4 we present Pearson's correlations[12] of the AESOP metrics with the Pyramid and Overall Responsiveness scores for TAC 2009 initial and update summaries excluding the model summaries (i.e., only system summaries).

[12] All with $p < 0.01$.

The layout of Table 4 is essentially the same as for Table 3 – top metrics on top three rows, baseline metrics on bottom two rows, and our metrics in between.

Interestingly, when the model summaries are excluded from the correlation analysis, the two baselines and three of our metrics yield increased correlation coefficients with respect to the case when these are included, whereas we note a significant drop for our *IC* metric. Thus, *IC* produced worse coefficients than the baselines (see italicised numbers).

Our best metric for initial summaries, pyramid score was *run4* and was ranked 2nd out of 37. Our best metric for update summaries, pyramid score was *run13* and was ranked 4th out of 37. Our best metric for initial summaries, overall responsiveness score was *IC+n-grams* and was ranked 3rd out of 37. Our best metric for update summaries, overall responsiveness score was *run4* and was ranked 6th out of 37.

Evaluation based on discriminative power. In order to evaluate metrics discriminative power, contingency tables with five distinctive cells were included, each containing the following information:

1. (Column1,Row1): pairs of summarizers (X,Y), where the AESOP metric and the Pyramid/Responsiveness method agree on both significant difference and polarity;
2. (Column2,Row1): pairs of summarizers (X,Y), where AESOP displays significance whereas Pyramid/Responsiveness method does not;
3. (Column1,Row2): pairs of summarizers (X,Y), where Pyramid/Responsiveness method displays significance whereas AESOP does not;
4. (Column2,Row2): pairs of summarizers (X,Y), where the AESOP metric and the Pyramid/Responsiveness method agree there is no significant difference;
5. (Column3,Row3): pairs of summarizers (X,Y), where Pyramid/Responsiveness method and AESOP agree there is significant difference but disagree on polarity.

In this analysis we propose to cast the discriminative power contingency tables in terms of precision, recall and balanced F1 measure based on the following criterion: if the AESOP metric displays significant differences then these cases are considered as ‘selected’ by the metric (i.e., result set), the rest are considered as ‘ignored’ by the metric. And vice versa, the significant differences according to the pyramid/responsiveness method are considered as the target set.

Table 5 shows the $P/R/F1$ measures capturing the discriminative power of our four AESOP metrics in the context of the two baselines (bottom two rows) and the top scoring metric(s) (top 2 rows for initial summaries, top row for update summaries) according to this evaluation criterion. Table 5 corresponds to the case where all summaries are included.

On the basis of the $F1$ measure alone, none of our metrics surpassed the baselines (see numbers in italics). In all cases the best $F1$ was yielded by our *run13*, which for the case of initial summaries, pyramid method ranked 15th out of 37, initial summaries, overall responsiveness method ranked 17th out of 37,

Run No.	Initial Summaries					
	Pyramid			Responsiveness		
	P	R	$F1$	P	R	$F1$
11,12,17,31,36	1.0	1.0	**1.0**	1.0	0.982	0.991
22				1.0	0.984	**0.992**
3	1.0	0.065	*0.123*	1.0	0.064	*0.112*
4	0.965	0.507	*0.665*	0.965	0.498	*0.657*
13	0.966	0.525	*0.681*	0.966	0.516	*0.673*
27	0.963	0.481	*0.642*	0.963	0.473	*0.634*
1	0.966	0.525	0.681	0.966	0.516	0.673
2	0.924	0.225	0.361	0.924	0.22	0.356
	Update Summaries					
11,12,17,31,36	1.0	0.998	**0.999**	1.0	0.984	**0.992**
3	1.0	0.005	0.009	1.0	0.005	0.009
4	0.952	0.319	0.478	0.945	0.312	0.469
13	0.964	0.427	0.592	0.958	0.419	0.583
27	0.947	0.286	0.44	0.94	0.28	0.432
1	0.964	0.427	0.592	0.958	0.419	0.583
2	0.934	0.229	0.367	0.925	0.223	0.36

Table 5. Discriminative power including models.

update summaries, pyramid method ranked 15th out of 37, update summaries, overall responsiveness method ranked 15th out of 37.

An interesting thing to note is that our *run3* in all cases consistently yielded a precision of 1.0, though on the expense of recall. This suggests a precise evaluation metric most likely suffering from data sparseness resulting from either terms not found in the MeSH taxonomy (a widely commented drawback of lexical databases such as MeSH and WordNet), or terms not seen in the training corpus (possibly less severe, since we used a basic back-off scheme by assigning default weights to such terms, though proper fine-tuning might further alleviate this problem).

In the case when no summary models are included the situation is similar; on $F1$ measure, none of our metrics surpassed the baselines.

Our best metric on overall responsiveness was *run27* and at initial summaries was ranked 6th, whereas at update summaries was ranked 7th out of 37. Our best metric on pyramid score was *run4* and at initial summaries was ranked 7th, whereas at update summaries was ranked 15th out of 37.

4 Conclusions

We discussed the methods used in our submissions to two TAC 2009 tasks: for multi-document summarisation (both basic and update), and automatic evaluation of summaries of peers.

Our systems did well at the basic summarisation task. Our *run19* was best in linguistic quality and second in overall responsiveness; the other run, *run11*, was second best in linguistic quality and seventh overall. In both runs LSA representations of all sentences were computed. In *run11*, the person, organisation and location names encountered in a sentence and automatically extracted using the EMM tools were used as terms in the representation of each sentence, in addition to word unigram and bigrams. The good results obtained suggest that named entity information improves sentence selection; we also observed that the fact that we only recognise full names (first and last name) slightly favours sentences from the beginning of the document, as such full names are more likely to occur early on in the document. In *run19*, in addition, MeSH thesaurus term mentions were included. Such information helped improve the results further, presumably because it provides more abstract content features than the words themselves. Interestingly, these good human evaluation results did not correspond to good automatic scores.

We conclude from the less outstanding results in the update summary task (positions 8 to 14 out of 52 submissions) that the same features may not be optimal for update summaries. Alternatively, our method to favour those features that are frequent in the new documents but not in the old documents may not be as successful as intuition tells us.

Regarding the approaches we submitted to automatically evaluate the summaries produced by peers, we did again better at overall responsiveness (positions 6 and 7 out of 37 submitted runs, for basic and for update summaries, respectively) than at other measures such as the pyramid score (positions 7 and 15 out of 37). Altogether, the results we achieved seem rather heterogeneous for this task, which makes it harder to draw conclusions.

As a summary, we conclude from the rather encouraging results in the basic summary task that LSA can in principle be a viable sentence selection method and that adding meta-information to the otherwise purely word-based representation of sentences is indeed very helpful. In future work, we intend to carry out experiments adding even more meta-information to the sentence representation, possibly exploiting the whole range of multilingual information extraction tools available to us. We believe furthermore that we should consider entity co-reference in our sentence representation.

Bibliography

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

Ani Nenkova and Rebecca Passonneau. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2004.

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

Bruno Pouliquen, Marco Kimler, Ralf Steinberger, Camelia Ignat, Tamara Oellinger, Ken Blackler, Flavio Fuart, Wajdi Zaghouani, Anna Widiger, Ann-Charlotte Forslund, and Clive Best. 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 2006.

Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, November 1995.

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

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

Josef Steinberger, Massimo Poesio, Mijail A. Kabadjov, and Karel Ježek. Two uses of anaphora resolution in summarization. *Information Processing and Management*, 43(6):1663–1680, 2007. Special Issue on Text Summarisation (Donna Harman, ed.).

Ralf Steinberger, Bruno Pouliquen, and Calemia Ignat. 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, 2009.