

Multilingual Statistical News Summarisation: Preliminary Experiments with English

Mijail Kabadjov*, Josef Steinberger†, Bruno Pouliquen*,
Ralf Steinberger* and Massimo Poesio‡§

*Joint Research Centre, European Commission, Italy; Email: *firstname.lastname@jrc.it*

† University of West Bohemia, Czech Republic; Email: *jstein@kiv.zcu.cz*

‡ University of Essex, United Kingdom

§ University of Trento, Italy; Email: *massimo.poesio@unitn.it*

Abstract—In this paper we present a generic approach for summarising multilingual news clusters such as the ones produced by the Europe Media Monitor (EMM) system. It is generic because it uses robust statistical techniques to perform the summarisation step and its multilinguality is inherited from the multilingual entity disambiguation system used to build the source representation. We ran preliminary experiments with the TAC 2008 data, an English corpus for summarisation research, and we obtained promising improvements over a summarisation system ranked in the top 20% at the TAC 2008 competition.

Keywords—Text Summarisation; Multilingual Text Mining; Entity Disambiguation; Latent Semantic Analysis;

I. INTRODUCTION

The Europe Media Monitor (EMM) news gathering engine collects over 80K news articles per day in about 50 languages from about 2.2K web news sources [1]. It feeds the news articles into four publicly accessible media monitoring applications (see <http://press.jrc.it/overview.html>), each with a different focus. The EMM applications cluster all these articles into major news stories, identify entities (locations, persons, organisations) [2], [3], send out breaking news alerts to subscribed users, monitor the development of a story over time, link news clusters across languages [4], and more. Currently, however, it does not provide succinct and comprehensive summaries for the news clusters. This is clearly a desirable feature since these clusters may contain hundreds of news articles which would otherwise be impossible to read in full within a short time frame and yet, this is often the case of decision makers within the European Union which make use of the EMM system on a daily or even hourly basis.

Another reason for needing to put a summarisation module in place within the EMM system is to bypass potential copyright issues. Currently, the EMM system serves as a news aggregator and general trend visualisation tool. However, for obvious reasons, it can only point to the source articles for further examination of a given news story. Providing a high quality summary on site would substantially improve the usability of EMM as a news aggregation and

trend visualisation system.

A distinctive characteristic of EMM that significantly adds to the overall complexity of the system is the high multilinguality of the raw data that the system must handle. Thus, one of the main challenges of an effort to develop a news summariser for the EMM system is that such a summariser must be necessarily multilingual, an issue that has not been addressed much in the vast literature on text summarisation.

In this work we address the problem of automatically producing summaries for the multilingual news clusters produced by the EMM system. The remainder of the paper is organised as follows: in section II we discuss the related work; in section III we present the Latent Semantic Analysis model for summarisation which we use as our starting point; in section IV we describe the system for multilingual entity disambiguation we used to enhance the LSA representation; in section V we put forward our proposal for the summarisation of multilingual news clusters; in section VI we present and discuss preliminary results over the TAC 2008 data and in the last section we conclude the paper with pointers to future work.

II. RELATED WORK

Work on Text Summarisation has been quite varied and abundant. A basic processing model for Text Summarisation, proposed by Sparck-Jones [5] 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 [6], cue phrases (e.g., “in conclusion”, “in summary”) and location (e.g., title, section

headings) [7]; there are machine learning approaches that combine a number of surface features [8] and/or more elaborate features exploiting discourse structure [9] to train classifiers using specialised 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 [10] and lexical cohesion [11] to identify salience or purely lexical approaches trying to identify ‘implicit topics’ by conflating together words using methods inspired by Latent Semantic Analysis LSA [12]; using yet deeper linguistic analysis, there are approaches purely based on discourse structure (e.g., RST) [13] and others combining discourse structure with surface features [14] or lexical with higher level semantic information such as anaphora [15]; and finally there are knowledge-rich approaches, where the source undergoes a substantial semantic analysis during the process of filling in a predefined template [16] or the source data is available in a more structured way (i.e., events have been identified already) [17].

III. LSA-BASED SUMMARISATION

Originally proposed by [12] and later improved by [18], this approach first builds a term-by-sentence matrix from the source, then applies a powerful statistical technique for matrix decomposition called Singular Value Decomposition (SVD) and finally uses the resulting matrices to identify and extract the most salient sentences. The idea behind this is that the SVD finds the latent (orthogonal) dimensions, which in simple terms correspond to the different topics discussed in the source (fine-tuning of the model is necessary, though, for optimal performance).

More formally, we first build matrix $A = [A_1 \dots A_n]$, where each column $A_i = [a_{1i} \dots a_{ni}]^T$ represents the weighted term-frequency vector of sentence i in a given document. Each element in this vector is defined as $a_{ji} = L(t_{ji}) \cdot G(t_{ji})$, where t_{ji} denotes the frequency with which term j occurs in sentence i , $L(t_{ji})$ is the local weight for term j in sentence i , and $G(t_{ji})$ is the global weight for term j in the whole document. We use a binary local weight and an entropy-based global weight (for more details see [19]).

Once the matrix A is built, it is decomposed via Singular Value Decomposition (SVD) defined as $A = U\Sigma V^T$. Due to space constraints, details of how sentences are extracted using matrices V^T and Σ are omitted here (see [19] for details on that).

IV. MULTILINGUAL ENTITY RECOGNITION AND DISAMBIGUATION

The disambiguation of entities in free text consists of first recognising a named entity in the text and then grounding it to an entity in the real world, say a location, a person or an organisation. The EMM system includes modules for entity recognition and disambiguation in 19 languages

[2], [3]. These modules are now being used as part of the summarisation task to add normalised and disambiguated structural information as input for LSA. In EMM, person, organisation and location entities are identified by unique language-independent identifiers which could be used as anchors to link related summaries across languages.

A. Recognition and Disambiguation of Geographical Locations

Historically (e.g., MUC-7 [20]), place name recognition consists of identifying references to locations in text and to disambiguate them from homographic person names or from other homographic words. For instance, there are places called ‘Javier’ (Spain) and ‘Solana’ (Philippines), and there are places called ‘And’ (Iran), ‘To’ (Ghana) and ‘Be’ (India). Within the EMM framework, [2] go beyond this MUC task by furthermore disambiguating between various homographic place names in order to identify precise latitude-longitude information and to put a dot on a map. For example, there are 15 different locations each with the names of ‘Berlin’, ‘Paris’ and ‘Roma’, and there are 205 places called ‘San Antonio’. In our experiments we make use of that EMM component to augment the term-by-sentence matrix (Fig. 1) with disambiguated and normalised location information.

B. Recognition and Disambiguation of Organisation and Person Names

We additionally make use of the multilingual person and organisation recognition tools described in [3]. What distinguishes this tool from others is its high multilinguality and, most of all, the functionality to map name variants referring to the same entity. Name spelling variation is not only a multilingual phenomenon, but it is even very frequent within a single language. [3] identified up to 170 ways of spelling the same name (all found in real news text). By recognising and mapping existing name variants for the same entity and by feeding this normalised information to the LSA representation (as described in Sec. V), additional useful cross-document links can be established.

V. EXTENDING LSA WITH SEMANTIC INFORMATION

In the work presented here, we propose to add normalised and disambiguated entity information extracted from the news texts to the LSA representation normally relying only on lexical information. The intuition is that words alone are weaker anchors due to morphological variance and synonymy, while the extracted entity information used has been disambiguated and normalised. For instance, person names are frequently written differently across documents [3], but after name variant recognition they can all be represented in the same way.

In our preliminary experiments, we followed the approach adopted by [15] and generalised the notion of ‘term’ to

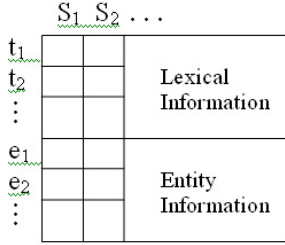


Figure 1. Augmented term-by-sentence matrix

Approach	R_1	R_2	R_{SU4}
lexical only TAC-08	0.355	0.088	0.123
lexical only	0.359	0.087	0.125
lexical+entities	0.367	0.093	0.13
Best TAC-08 system		0.111	0.142

Table I
MULTI-DOCUMENT SUMMARISATION RESULTS

entail not only word tokens, but also references to real entities. Thus, we likewise extended the term-by-sentence matrix A used as input to the LSA with information about disambiguated entities (see Fig. 1). However, as opposed to [15] we did not use a coreference resolver.¹ Instead, we used a more general multilingual named entity disambiguator and geo-tagger (cf. previous section).

Augmenting the initial matrix with information about disambiguated entities naturally provides not only stronger inter-sentential cohesion (i.e., the LSA clusters sentences from different documents that make reference to the same entities), but also provides multilingual capabilities inherited by the multilingual entity disambiguation. Thus, this approach to summarisation is not only multi-document, but also multilingual.²

VI. PRELIMINARY RESULTS

Using a standard English corpus for Summarisation research developed by the US National Institute for Standards and Technology (NIST) for the 2008 Text Analysis Conference (hereafter TAC-08), we obtained promising, though, not statistically significant improvements over a lexical-only baseline ranked in the top 15%-24% across all evaluation metrics at the 2008 TAC competition. For our preliminary experiments we used the popular ROUGE metric to evaluate the performance. The results are presented in Tables I, II and III.

On the standard multi-document summarisation task (see table I), we include the official scores at TAC-08 of a lexical-only summariser that we used as a baseline for our

¹Steinberger et al. 2007 worked on monolingual single-document summarisation.

²The multilingual named entity disambiguator and geo-tagger developed at the JRC have already been used for cross-lingual linking of multilingual news clusters produced by the EMM system [4].

Approach	R_1	R_2	R_{SU4}
TAC-08	0.348	0.081	0.12
lexical only	0.363	0.091	0.13
lexical+entities	0.364	0.92	0.132
Best TAC-08 system		0.101	0.136

Table II
UPDATE SUMMARISATION RESULTS

Approach	R_1	R_2	R_{SU4}
TAC-08	0.352	0.084	0.122
lexical only	0.361	0.089	0.128
lexical+entities	0.366	0.093	0.131

Table III
OVERALL PERFORMANCE RESULTS

experiments (cf. first row of the table) as well as an improved version of it referred to as ‘lexical only’ (cf. second row). The third row of table I corresponds to the results obtained by the LSA extension proposed in section V. The last row shows the results obtained by the best system at TAC-08 and is included only for reference purposes.

From Table I we can see that the performance of the ‘lexical+entities’ version of the system is higher than the ‘lexical only’ version, our baseline, but we note the improvement is not statistically significant after running a t test.³

It is worth noting that the EMM entity disambiguation module we used in this experiment has been optimised for precision, since in the EMM’s context the vast amounts of data (i.e., over 80K processed articles per day) make up for the compromise on recall. However, in the TAC-08 context there is substantial room for improvement of the recall of entity mentions within a document by bringing in an intra-document coreference resolution system, such as GuiTAR [21], and aggregating the output with that of the entity disambiguator. In the light of this we believe the attained results are promising.

On the update summarisation task (see table II),⁴ we present the same evaluation dimensions as for the standard summarisation task. The picture is similar to the previous case, though, the improvement this time is clearly insignificant. We believe this is possibly due to this second task being much more specific than the standard summarisation task and hence needing either more elaborate fine-tuning of the model or a much bigger corpus for a larger-scale evaluation.

Table III presents a combined picture of both tasks.

Currently we are in the process of error analysis to

³Note that the statistical test we used to attest significance was ran against the improved version of the lexical-only summariser and not the official TAC-08 scores, since we considered it was the fairer comparison.

⁴The purpose of the update summarisation task is to produce a summary of only the novel information contained in a newer set of news articles with respect to an older set, both covering the same news story.

identify the major sources of errors.

VII. CONCLUSION

We presented a generic statistical approach to summarisation based on the Latent Semantic Analysis paradigm. The key extension to previous LSA-based work is the use of a multilingual and multi-document entity disambiguation system when building the source representation. The entity disambiguation system is able to identify and ground references to geographical locations in news articles in 19 languages to the corresponding entries in gazetteers as well as to link references to persons and organisations to respective records in a continuously updated database of known names built over the past eight years. We showed promising results from preliminary experiments with English using the Text Analysis Conferences corpus from 2008.

Next steps will include experiments such as adding intra-document co-reference (e.g., Obama, Barack Obama and US President) and adding further normalised entity information (e.g., dates, vehicles, professions and nationality expressions). At the workshop, we will show examples of multilingual summaries automatically produced by the system.

ACKNOWLEDGMENT

We would like to thank the EMM team for providing such a stable and robust media monitoring infrastructure.

REFERENCES

- [1] M. Atkinson and E. V. der Goot, "Near real time information mining in multilingual news," in *Proceedings of the 18th International World Wide Web Conference (WWW 2009)*, (Madrid, Spain), pp. 1153–1154, April 2009.
- [2] B. Poulliquen, M. Kimler, R. Steinberger, C. Ignat, T. Oellinger, K. Blackler, F. Fuart, W. Zaghouni, A. Widiger, A.-C. Forslund, and C. Best, "Geocoding multilingual texts: Recognition, disambiguation and visualisation," in *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, (Genoa, Italy), pp. 53–58, May 2006.
- [3] B. Poulliquen and R. Steinberger, "Automatic construction of multilingual name dictionaries," in *Learning Machine Translation* (C. Goutte, N. Cancedda, M. Dymetman, and G. Foster, eds.), MIT Press, NIPS series, 2009.
- [4] R. Steinberger, B. Poulliquen, and C. Ignat, "Using language-independent rules to achieve high multilinguality in text mining," in *Mining Massive Data Sets for Security* (F. Fogelman-Soulié, D. Perrotta, J. Piskorski, and R. Steinberger, eds.), Amsterdam, Holland: IOS-Press, 2009.
- [5] K. S. Jones, "Automatic summarising: Factors and directions," in Mani and Maybury [22].
- [6] H. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development*, vol. 2, no. 2, pp. 159–165, 1958.
- [7] H. Edmundson, "New methods in automatic extracting," *Journal of the Association for Computing Machinery*, vol. 16, no. 2, pp. 264–285, 1969.
- [8] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer," in *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (Seattle, Washington), pp. 68–73, 1995.
- [9] S. Teufel and M. Moens, "Sentence extraction as a classification task," in Mani [23].
- [10] B. Boguraev and C. Kennedy, "Salience-based content characterisation of text documents," in Mani [23].
- [11] R. Barzilay and M. Elhadad, "Using lexical chains for text summarization," in Mani [23].
- [12] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in *Proceedings of ACM SIGIR*, (New Orleans, US), 2002.
- [13] D. Marcu, "From discourse structures to text summaries," in Mani [23].
- [14] E. Hovy and C. Lin, "Automated text summarization in summarist," in Mani [23].
- [15] J. Steinberger, M. Poesio, M. A. Kabadjov, and K. Ježek, "Two uses of anaphora resolution in summarization," *Information Processing and Management*, vol. 43, no. 6, pp. 1663–1680, 2007. Special Issue on Text Summarisation (Donna Harman, ed.).
- [16] K. McKeown and D. Radev, "Generating summaries of multiple news articles," in *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (Seattle, Washington), pp. 74–82, 1995.
- [17] M. Maybury, "Generating summaries from event data," in Mani and Maybury [22].
- [18] J. Steinberger and K. Ježek, "Text summarization and singular value decomposition," in *Proceedings of the 3rd ADVIS conference*, (Izmir, Turkey), 2004.
- [19] J. Steinberger and K. Ježek, "Sutler: Update summarizer based on latent topics," in *Text Analysis Conference*, 2008.
- [20] L. Hirschman, "MUC-7 coreference task definition, version 3.0," in *Proceedings of the 7th Message Understanding Conference* (N. Chinchor, ed.), NIST, 1998. Available online at http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_toc.html.
- [21] M. A. Kabadjov, *A Comprehensive Evaluation of Anaphora Resolution and Discourse-new Recognition*. PhD thesis, Department of Computer Science, University of Essex, December 2007.
- [22] I. Mani and M. Maybury, eds., *Advances in Automatic Text Summarization*, MIT Press, 1999.
- [23] I. Mani, ed., *Proceedings of the Workshop on Intelligent and Scalable Text Summarization at the Annual Joint Meeting of the ACL/EACL*, (Madrid), 1997.