Content Selection through Paraphrase Detection: Capturing different Semantic Realisations of the Same Idea

Elena Lloret
University of Alicante
Alicante, Spain
elloret@dlsi.ua.es

Claire Gardent
CNRS/LORIA
Nancy, France
claire.gardent@loria.fr

1 Introduction

Summarisation can be seen as an instance of Natural Language Generation (NLG), where *"what to say"* corresponds to the identification of relevant information, and *"how to say it"* would be associated to the final creation of the summary. When dealing with data coming from the Semantic Web (e.g., RDF triples), the challenge of how a good summary can be produced arises. For instance, having the RDF properties from an infobox of a Wikipedia page, how could a summary expressed in natural language text be generated? and how could this summary sound as natural as possible (i.e., be an abstractive summary) far from only being a bunch of selected sentences output together (i.e., extractive summary)? This would imply to be able to successfully map the RDF information to a semantic representation of natural language sentences (e.g., predicate-argument (*pred-arg*) structures). Towards the long-term objective of generating abstractive summaries from Semantic Web data, the specific goal of this paper is to propose and validate an approach to map linguistic structures that can encode the same meaning but with different words (e.g., sentence-to-sentence, *pred-arg*-to-*pred-arg*, RDF-to-TEXT) using continuous semantic representation of text. The idea is to decide the level of document representation to work with; convert the text into that representation; and perform a pairwise comparison to decide to what extent two pairs can be mapped or not. For achieving this, different methods were analysed, including traditional Wordnet-based ones, as well as more recent ones based on word embeddings. Our approach was tested and validated in the context of document-abstract sentence mapping to check whether it was appropriate for identifying important information. The results obtained good performance, thus indicating that we can rely on the approach and apply it to further contexts (e.g., mapping RDFs into natural language).

The remainder of this paper is organised as follows: Section 2 outlines related work. Section 3 explains the proposed approach for mapping linguistic units. Section 4 describes our dataset and experiments. Section 5 provides the results and discussion. Finally, Section 6 draws the main conclusions and highlights possible futures directions.

2 Related Work

Abstractive summarisation is one of the most challenging issues to address automatically, since it both requires deep language understanding and generation with a strong semantic component. For tackling this task, approaches usually need to define an internal representation of the text, that can be in the form of SVO triples (Genest and Lapalme, 2011), basic semantic units consisting of *actor-action-receiver* (Li, 2015), or using *pred-arg* structures (Khan et al., 2015). In this latter work, *pred-arg* structures extracted from different related documents are compared, so that common or redundant information can be grouped into clusters. For computing a similarity matrix, Wordnet[1]-based similarity metrics are used, mainly relying on the semantic distance between concepts, given Wordnets' hierarchy.

On the other hand, previous works on linguistic structure mapping can be related to paraphrase identification (Fernando and Stevenson, 2008; Xu et al., 2015), as well as to *pred-arg* alignment (Wolfe et al., 2015; Roth and Frank, 2015). However, these works only use semantic similarity metrics based on WordNet or other semantic resources, such as ConceptNet[2] or FrameNet[3].

[1] https://wordnet.princeton.edu/
[2] http://conceptnet5.media.mit.edu/
[3] https://framenet.icsi.berkeley.edu/fndrupal/

The use of continuous semantic representation, and in particular the learning or use of Word Embeddings (WE) has been shown to be more appropriate and powerful approach for representing linguistic elements (words, sentences, paragraphs or documents) (Turian et al., 2010; Dai et al., 2015). Given its good performance, they have been recently applied to many natural language generation tasks (Collobert et al., 2011; Kågebäck et al., 2014). The work presented in (Perez-Beltrachini and Gardent, 2016) proposes a method to learn embeddings to lexicalise RDF properties, showing also the potential of using this type of representation for the Semantic Web.

3 Our Mapping Approach

Our approach mainly consists of three stages: i) identification and extraction of text semantic structures; ii) representation of these semantic structures in a continuous vector space; and iii) define and compute the similarity between two representations.

For the first stage, depending on the level defined for the linguistic elements (e.g., a clause, a sentence, a paragraph), a text processing is carried out, using the appropriate tools to obtain the desired structures (e.g., sentence segmentation, semantic role labelling, syntactic parsing, etc.). Then, in the second stage, we represent each structure through its WEs. If the structure consists of more than one element, we will compute the final vector as the composition of the WEs of each of the elements it contains. This is a common strategy that has been previously adopted, in which the addition or product normally lead to the best results (Mitchell and Lapata, 2008; Blacoe and Lapata, 2012; Kågebäck et al., 2014). Finally, the aim of the third stage is to define a similarity metric between the vectors obtained in the second stage.

4 Dataset and Approach Configuration

The English training collection of documents and abstracts from the Single document Summarization task (MSS)[4] of the MultiLing2015 was used as corpus. It consisted of 30 Wikipedia documents from heterogeneous topics (e.g., history of Texas University, fauna of Australia, or Magic Johnson) and their abstracts, which corresponded to the introductory paragraphs of the Wikipedia page. Documents were rather long, having 3,972 words on average (the longest document had 8,348 words and the shortest 2,091), whereas abstracts were 274 words on average (the maximum value was 305 words and the minimum 243), thus resulting in a very low compression ratio[5] - around 7%.

For carrying out the experiments, our approach receives document-abstract pairs as input. These correspond to the source documents, as well as the abstracts associated to those documents. Following the stages defined in Section 3, both were segmented in sentences, and the *pred-arg* structures were automatically identified using SENNA semantic role labeller [6]. Different configurations were tested as far as the WE and the similarity metrics were concerned for the second and third stages. For representing either sentences or *pred-arg* structures, GLoVe pre-trained WE vectors (Pennington et al., 2014) were used, specifically the ones derived from Wikipedia 2014 + Gigaword 5 corpora, containing around 6 billion tokens; and the ones derived from a Common Crawl, with 840 billion tokens. Regarding the similarity metrics, Wordnet-based metrics included the shortest path between synsets, Leacock-Chodorow similarity, Wu-Palmer similarity, Resnik similarity, Jiang-Conrath similarity, and Lin similarity, all of them implemented in NLTK[7]. For the WE settings, the similarity metrics were computed on the basis of the cosine similarity and the Euclidean distance. These latter metrics were applied upon the two composition methods for sentence embedding representations: addition and product, as described in (Blacoe and Lapata, 2012). In the end, a total of 38 distinct configurations were obtained.

5 Evaluation and Discussion

We addressed the validation of the source document-abstract pairs mapping as an extrinsic task using ROUGE (Lin, 2004). ROUGE is a well-known tool employed for summarisation evaluation, which computes the n-gram overlapping between an automatic and a reference summary in terms of n-grams (unigrams - ROUGE 1; bigrams - ROUGE 2, etc.). Our assumption behind this type of evaluation was that considering the

[4] http://multiling.iit.demokritos.gr/pages/view/1532/task-mss-single-document-summarization-data-and-information

[5] The compression ratio is the size of the summary with respect to the source document, i.e., the percentage of relevant information to be kept.

[6] http://ronan.collobert.com/senna/

[7] http://www.nltk.org/

	ROUGE-1			ROUGE-2			ROUGE-SU4		
	R	P	F	R	P	F	R	P	F
TEXT baseline	41.63	40.64	41.11	10.11	9.87	9.99	15.67	15.29	15.45
Best TEXT+ WORDNET	42.04	41.58	41.78	11.40	11.25	11.32	16.55	16.34	16.43
Best TEXT+ WE	**50.36**	**47.99**	**49.12**	**17.35**	**16.56**	**16.94**	**22.51**	**21.46**	**21.96**
PRED-ARG baseline	34.64	34.05	34.24	7.19	7.09	7.12	12.25	12.04	12.10
Best PRED-ARG + WORDNET	38.45	38.45	38.39	9.97	9.98	9.96	14.79	14.80	14.77
Best PRED-ARG + WE	**46.88**	**45.17**	**45.97**	**15.18**	**14.53**	**14.84**	**20.02**	**19.23**	**15.60**

Table 1: Results (in percentages) for the extrinsic validation of the mapping.

source document snippets of the top-ranked mapping pairs, and directly building a summary with them (i.e., an extractive summary), good ROUGE results should be obtained if the mapping was good enough.

Table 1 reports the most relevant results obtained. As baselines, we considered the ROUGE direct comparison between the sentences (or *pred-arg* structures) of the source document and the ones in the abstract (TEXT baseline, and PRED-ARG baseline, respectively). We report the results for ROUGE-1, ROUGE-2 and ROUGE-SU4[8]. The results obtained show that representing the semantics of a sentence or *pred-arg* structure using WE leads to the best results, improving those from traditional WordNet-based similarity metrics. The best approach for the WE configuration corresponds to the addition composition method with cosine similarity, and using the pretrained WE derived from Wikipedia+GigaWord. Compared to the state of the art in summarisation, the results with WE are also encouraging, since previous published results with the same corpus (Alcón and Lloret, 2015) are close to 44% (F-measure for ROUGE-1).

Concerning the comparison between whether using the whole text with respect to only using the *pred-arg* structures, the former gets better results. This is logical since the more text to compare, the higher chances to obtain similar n-grams when evaluating with ROUGE. However, this also limits the capability of abstractive summarisation systems, since we would end up with selecting the sentences as they are, thus restricting the method to purely extractive. Nevertheless, the results obtained by the use of *pred-arg* structures are still reasonably acceptable, and this type of structure would allow to generalise the key content to be selected that should be later rephrased in a proper sentence, producing an abstractive summary. Next, we provide the top 3 best pair alignments (source document— abstract) of the highest performing configuration using *pred-arg* structure as examples. The value in brackets mean the similarity percentage obtained by our approach.

protected areas — protected areas (100%)
the insects comprising 75% of Australia 's known species of animals —The fauna of Australia consists of a huge variety of strange and unique animals ; some 83% of mammals, 89% of reptiles, 90% of fish and insects (99.94%)
European settlement , direct exploitation of native faun , habitat destruction and the introduction of exotic predators and competitive herbivores led to the extinction of some 27 mammal, 23 bird and 4 frog species. — Hunting, the introduction of non- native species, and land - management practices involving the modification or destruction of habitats led to numerous extinctions (99.93%)

Finally, our intuition behind the results obtained (maximum values of 50%) is that not all the information in the abstract can be mapped with the information of the source document, indicating that a proper abstract may contain extra information that provides from the world knowledge of its author.

6 Conclusion and Future Work

This paper presented an approach to automatically map linguistic structures using continuous semantic representation of sentences. The analysis conducted over a wide set of configuration showed that the use of WEs improves the results compared to traditional WordNet-based metrics, thus being suitable to be employed in data-to-text NLG approaches that need to align content from the Semantic Web to text in natural language. As future work, we plan to evaluate the approach intrinsically and apply it to map non-linguistic in-

[8]ROUGE-SU4 accounts for skip-bigrams with maximum gap length of 4.

formation (e.g., RDF) to natural language. We would also like to use the proposed method to create training positive and negative instances to learn classification models for content selection.

Acknowledgments

This research work has been partially funded by the University of Alicante, Generalitat Valenciana, Spanish Government and the European Commission through the projects, "Explotación y tratamiento de la información disponible en Internet para la anotación y generación de textos adaptados al usuario" (GRE13-15) and "DIIM2.0: Desarrollo de técnicas Inteligentes e Interactivas de Minería y generación de información sobre la web 2.0" (PROMETEOII/2014/001), TIN2015-65100-R, TIN2015-65136-C2-2-R, and SAM (FP7-611312), respectively.

References

Óscar Alcón and Elena Lloret. 2015. Estudio de la influencia de incorporar conocimiento léxico-semántico a la técnica de análisis de componentes principales para la generación de resúmenes multilingües. *Linguamatica*, 7(1):53–63, July.

William Blacoe and Mirella Lapata. 2012. A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 546–556, Jeju Island, Korea, July. Association for Computational Linguistics.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, November.

Andrew M. Dai, Christopher Olah, and Quoc V. Le. 2015. Document embedding with paragraph vectors. *CoRR*, abs/1507.07998.

Samuel Fernando and Mark Stevenson. 2008. A semantic similarity approach to paraphrase detection. *Computational Linguistics UK (CLUK 2008) 11th Annual Research Colloqium*.

Pierre-Etienne Genest and Guy Lapalme. 2011. Framework for abstractive summarization using text-to-text generation. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, MTTG '11, pages 64–73, Stroudsburg, PA, USA. Association for Computational Linguistics.

Atif Khan, Naomie Salim, and Yogan Jaya Kumar. 2015. A framework for multi-document abstractive summarization based on semantic role labelling. *Appl. Soft Comput.*, 30(C):737–747, May.

Mikael Kågebäck, Olof Mogren, Nina Tahmasebi, and Devdatt Dubhashi. 2014. Extractive summarization using continuous vector space models. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pages 31–39, Gothenburg, Sweden, April. Association for Computational Linguistics.

Wei Li. 2015. Abstractive multi-document summarization with semantic information extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1908–1913, Lisbon, Portugal, September. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio, June. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Laura Perez-Beltrachini and Claire Gardent. 2016. Learning Embeddings to lexicalise RDF Properties. In **SEM 2016: The Fifth Joint Conference on Lexical and Computational Semantics*, Berlin, Germany.

Michael Roth and Anette Frank. 2015. Inducing Implicit Arguments from Comparable Texts: A Framework and its Applications. *Computational Linguistics*, 41:625–664.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 384–394, Stroudsburg, PA, USA. Association for Computational Linguistics.

Travis Wolfe, Mark Dredze, and Benjamin Van Durme. 2015. Predicate argument alignment using a global coherence model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11–20, Denver, Colorado, May–June. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Bill Dolan. 2015. Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 1–11, Denver, Colorado, June. Association for Computational Linguistics.