Generating sets of related sentences from input seed features

Cristina Barros
Department of Software
and Computing Systems
University of Alicante
Apdo. de Correos 99
E-03080, Alicante, Spain
cbarros@dlsi.ua.es

Elena Lloret
Department of Software
and Computing Systems
University of Alicante
Apdo. de Correos 99
E-03080, Alicante, Spain
elloret@dlsi.ua.es

1 Introduction

The Semantic Web (SW) can provide Natural Language Generation (NLG) with technologies capable to facilitate access to structured Web data. This type of data can be useful to this research area, which aims to automatically produce human utterances, in its different subtasks, such as in the content selection or its structure.

NLG has been widely applied to several fields, for instance to the generation of recommendations (Lim-Cheng et al., 2014). However, generation systems are currently designed for very specific domains (Ramos-Soto et al., 2015) and predefined purposes (Ge et al., 2015). The use of SW's technologies can facilitate the development of more flexible and domain independent systems, that could be adapted to the target audience or purposes, which would considerably advance the state of the art in NLG. The main objective of this paper is to propose a multidomain and multilingual statistical approach focused on the surface realisation stage using factored language models. Our proposed approach will be tested in the context of two different domains (fairy tales and movie reviews) and for the English and Spanish languages, in order to show its appropriateness to different non-related scenarios. The main novelty studied in this approach is the generation of related sentences (sentences with related topics) for different domains, with the aim to achieve cohesion between sentences and move forward towards the generation of coherent and cohesive texts. The approach can be flexible enough thanks to the use of an input seed feature that guides all the generation process. Within our scope, the seed feature can be seen as an abstract object that will determine how the sentence will be in terms of content. For example, this seed feature could be a phoneme, a property or a RDF triple from where the proposed approach could generate a sentence.

2 Factored Language Models and NLG

Factored language models (FLM) are an extension of language models proposed in (Bilmes and Kirchhoff, 2003). In this model, a word is viewed as a vector of *k* factors such that $w_t \equiv \{f_t^1, f_t^2, \ldots, f_t^K\}$. These factors can be anything, including the Part-Of-Speech (POS) tag, lemma, stem or any other lexical, syntactic or semantic feature. Once a set of factors is selected, the main objective of a FLM is to create a statistical model $P(f|f_1, \ldots, f_N)$ where the prediction of a feature f is based on *N* parents $\{f_1, \ldots, f_N\}$. For example, if *w* represents a word token and *t* represents a POS tag, the expression $P(w_i|w_{i-2}, w_{i-1}, t_{i-1})$ provides a model to determine the current word token, based on a traditional n-gram model together with the POS tag of the previous word. Therefore, in the development of such models there are two main issues to consider: 1) choose an appropriate set of factors, and 2) find the best statistical model over these factors.

In recent years, FLM have been used in several areas of Computational Linguistics, mostly in machine translation (Crego, 2010; Axelrod, 2006) and speech recognition (Tachbelie et al., 2011; Vergyri et al., 2004). To a lesser extent, they have been also employed for generating language, mainly in English. This is the case of the BAGEL system (Mairesse and Young, 2014), where FLM (with semantic concepts as factors) are used to predict the semantic structure of the sentence that is going to be generated; or OpenCCG (White and Rajkumar, 2009), a surface realisation tool, where FLM (with POS tag and supertags as factors) are used to score partial and complete realisations to be later selected. More recently, FLM (with POS tag, word and lemma as factors) were used to

rank generated sentences in Portuguese (Novais and Paraboni, 2012).

The fact of generating connected and related sentences is a challenge in itself, and, to the best of our knowledge there is not any research with the restriction of containing words with a specific seed feature, thus leading to a more flexible NLG approach that could be easily adapted to different purposes, domains and languages.

3 Generating Related Sentences Using FLM

We propose an almost-fully language independent statistical approach focused on the surface realisation stage and based on over-generation and ranking techniques, which can generate related sentences for different domains. This is achieved through the use of input seed features, which are abstract objects (e.g., a phoneme, a semantic class, a domain, a topic, or a RDF triple) that will guide the generation process in relation to the most suitable vocabulary for a given purpose or domain.

Starting from a training corpus, a test corpus and a seed feature as the input of our approach, a FLM will be learnt over the training corpus and a bag of words with words related with the seed feature will be obtained from the test corpus. Then, based on the FLM and bag of words previously obtained, the process will generate several sentences for a given seed feature, which will be subsequently ranked. This process will prioritise the selection of words from the bag of words to guarantee that the generated sentences will contain the maximum number of words related with the input seed feature. Once several sentences are generated, only one of them will be selected based on the sentence probability, that will be computed using a linear combination of FLMs.

When a sentence is generated, we will perform post-tagging, syntactic parsing and/or semantic parsing to identify several linguistic components of the sentence (such as the subject, named entities, etc.) that will also provide clues about its structural shape. This will allow us to generate the next sentence taking into account the shape of the previous generated one, and the structure we want to obtain (e.g., generating sentences about the same subject with complementary information).

4 Experimental scenarios and resources

For our experimentation, we want to consider two different scenarios, NLG for assistive technologies and sentiment-based NLG. Within the first scenario, the experimentation will be focused on the domain of fairy tales. The purpose in this scenario is the generation of stories that can be useful for therapies in dyslalia speech therapies (Rvachew et al., 1999). Dyslalia is a disorder in phoneme articulation, so the repetition of words with problematic phonemes can improve their pronunciation. Therefore, in this scenario, the selected seed feature will be a phoneme, where the generated sentences will contain a large number of words with a concrete phoneme. As corpora, a collection of Hans Christian Andersen tales will be used due to the fact that its vocabulary is suitable for young audience, since dyslalia affects more to the child population, having a 5-10% incidence among them (Conde-Guzón et al., 2014).

Regarding the second scenario, the experimentation will be focused on generating opinionated sentences (i.e., sentences with a positive or negative polarity) in the domain of movie reviews. Taking into account that there are many Websites where users express their opinions by means of non-linguistic rates in the form of numeric values or symbols[1], the generation of this kind of sentences can be used to generate sentences from visual numeric rates. Given this proposed scenario, we will employ the Spanish Movie Reviews corpus[2] and the Sentiment Polarity Dataset (Pang and Lee, 2004) as our corpora for Spanish and English, respectively.

In order to learn the FLM that will be used during the generation, we will use SRILM (Stolcke, 2002), a software which allows to build and apply statistical language models, which also includes an implementation of FLM.

In addition, Freeling language analyser (Padró and Stanilovsky, 2012) will be also employed to tag the corpus with lexical information as well as to perform the syntactic analysis and the name entity recognition of the generated sentences. Furthermore, in order to obtain and evaluate the polarity for our second proposed domain, we will employ the sentiment analysis classifier described and developed in (Fernández et al., 2013).

[1] An example of such a Website can be found at: http://www.reviewsdepeliculas.com/

[2] http://www.lsi.us.es/ fermin/corpusCine.zip

5 Preliminary Experimentation

As an initial experimentation, we design a simple grammar (based on the basic clause structure that divides a sentence into subject and predicate) to generate sets of sentences which will have related topics (nouns) with each other, since these topics will appear within the set.

In this case, we generate the sentences with the structure shown in Figure 1, where we use the direct object of the previous generated sentences as the subject for the following sentence to be produced, so that we can obtain a preliminary set of related sentences.

The words contained in these preliminary related sentences are in a lemma form since this configuration proved to works better than others, being able to be further inflected in order to obtain several inflections of the sentences from where the final generated one will be chosen.

$$S \rightarrow NP\ VP$$
$$NP \rightarrow D\ N$$
$$VP \rightarrow V\ NP$$

Figure 1: Basic clause structure grammar.

With this structure we generated a set of 3 related sentences for each phoneme in both languages, Spanish and English, and another set of 3 related sentences for positive and negative polarities in the languages mentioned before.

These sentences have the structure seen above and were ranked according to the approach outlined in section 3 being the linear combination of FLM as follows: $P(w_i) = \lambda_1 P(f_i|f_{i-2}, f_{i-1}) + \lambda_2 P(f_i|p_{i-2}, p_{i-1}) + \lambda_3 P(p_i|f_{i-2}, f_{i-1})$, where f can be can be either a lemma and a word, p refers to a POS tag, and λ_i are set $\lambda_1 = 0.25$, $\lambda_2 = 0.25$ and $\lambda_3 = 0.5$. These values were empirically determined.

Some examples of the generated sentences for the first scenario, concerning the generation of sentences for assistive technologies, is shown in Figure 2. In some of the sets of generated sentences, the same noun appears as a direct object in both, the first and the third generated sentences for that set. On the other hand, examples of sets of sentences generated in both, English and Spanish, for the second experimentation scenario (movie reviews domain) are shown in the Figure 3.

Generally, the generated sentences for our two experimentation scenarios, conform to the specified in section 4, although in some cases the verbs in these sentences need the inclusion of preposition in order to bring more correctness to the generated sentences.

Spanish
Phoneme: */n/*
Cuánto cosa tener nuestro pensamiento.
(How much thing have our thinking.)
Cuánto pensamiento tener nuestro corazón.
(How much thought have our heart.)
Cuánto corazón tener nuestro pensamiento.
(How much heart have our thinking.)

English
Phoneme: */s/*
These child say the princess.
Each princess say the shadow.
Each shadow pass this story.

Figure 2: Example generated sentences for the assistive technologies scenario.

Spanish
Polarity: *Negative*
Este defecto ser el asesino.
(This defect being the murderer.)
Su asesino ser el policía.
(His murderer be the police.)
El policía interpretar este papel.
(The police play this role.)

English
Polarity: *Negative*
Many critic reject the plot.
This plot confuse the problem.
The problem lie this mess.

Figure 3: Example generated sentences for movie reviews domain in our second scenario.

At this stage, these preliminary set of generated related sentences are a promising step towards our final goal, since the number of words with the seed feature among the sentences are more than the number of words of the sentences, meeting the overall objective for which they were generated. Although the grammar used in the generation of these sentences only captures the basic structure for the two languages studied, the use of more complex grammars could give us insights to improve some aspects of the generation of these preliminary sentences in the future.

6 Ongoing research steps

In order to enrich this approach and meet the final goal, we want to deeply research into some of the representation languages used by the SW, such as OWL, as well as its technologies, that fit our proposed approach. Obtaining information related to a certain topic is tough without using any kind of

external technology, so the employing of SW languages, such as RDF, can facilitate us accessing this type of information.

In the future, we would like to analyse how the generated sentences could be connected using discourse markers. We also would like to test the generation of sentences using other structural shapes, such as sharing the same subject or sentences sharing the same predicative objects with different subjects. The generation of related sentences is not a trivial task, being the cohesion and coherence between sentences very hard to be checked automatically. So, in that case, we plan to conduct an exhaustive user evaluation of the generated sentences using crowdsourcing platforms.

Acknowledgments

This research work has been funded by the University of Alicante, Generalitat Valenciana, Spanish Government and the European Commission through the projects GRE13-15, PROMETEOII/2014/001, TIN2015-65100-R, TIN2015-65136-C2-2-R, and FP7-611312, respectively.

References

Amittai Axelrod. 2006. Factored language models for statistical machine translation. master thesis. university of edinburgh.

Jeff A. Bilmes and Katrin Kirchhoff. 2003. Factored language models and generalized parallel backoff. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume of the Proceedings of HLT-NAACL 2003–short Papers - Volume 2*, pages 4–6.

Pablo Conde-Guzón, Pilar Quirós-Expósito, María Jesús Conde-Guzón, and María Teresa Bartolomé-Albistegui. 2014. Perfil neuropsicológico de niños con dislalias: alteraciones mnésicas y atencionales. *Anales de Psicología*, 30:1105 – 1114.

François Crego, Josep M.and Yvon. 2010. Factored bilingual n-gram language models for statistical machine translation. *Machine Translation*, 24(2):159–175.

Javi Fernández, Yoan Gutiérrez, José Manuel Gómez, Patricio Martínez-Barco, Andrés Montoyo, and Rafael Muñoz. 2013. Sentiment analysis of spanish tweets using a ranking algorithm and skipgrams. *Proc. of the TASS workshop at SEPLN 2013*, pages 133–142.

Tao Ge, Wenzhe Pei, Heng Ji, Sujian Li, Baobao Chang, and Zhifang Sui. 2015. Bring you to the past: Automatic generation of topically relevant event chronicles. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 575–585, July.

Natalie R. Lim-Cheng, Gabriel Isidro G. Fabia, Marco Emil G. Quebral, and Miguelito T. Yu. 2014. Shed: An online diet counselling system. In *DLSU Research Congress 2014*.

François Mairesse and Steve Young. 2014. Stochastic language generation in dialogue using factored language models. *Comput. Linguist.*, 40(4):763–799.

Eder Miranda Novais and Ivandré Paraboni. 2012. Portuguese text generation using factored language models. *Journal of the Brazilian Computer Society*, 19(2):135–146.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*.

A. Ramos-Soto, A. J. Bugarn, S. Barro, and J. Taboada. 2015. Linguistic descriptions for automatic generation of textual short-term weather forecasts on real prediction data. *IEEE Transactions on Fuzzy Systems*, 23(1):44–57.

Susan Rvachew, Susan Rafaat, and Monique Martin. 1999. Stimulability, speech perception skills, and the treatment of phonological disorders. *American Journal of Speech-Language Pathology*, 8(1):33–43.

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing, vol 2.*, pages 901–904.

Martha Yifiru Tachbelie, Solomon Teferra Abate, and Wolfgang Menzel, 2011. *Human Language Technology. Challenges for Computer Science and Linguistics: 4th Language and Technology Conference*, chapter Morpheme-Based and Factored Language Modeling for Amharic Speech Recognition, pages 82–93. Springer Berlin Heidelberg.

Dimitra Vergyri, Katrin Kirchhoff, Kevin Duh, and Andreas Stolcke. 2004. Morphology-based language modeling for arabic speech recognition. In *INTERSPEECH*, volume 4, pages 2245–2248.

Michael White and Rajakrishnan Rajkumar. 2009. Perceptron reranking for ccg realization. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 410–419. Association for Computational Linguistics.