Controlled generation of synthetic corpora for NLP evaluation

Jérémie Démarchez, Cyril Labbé

▶ **To cite this version:**

Jérémie Démarchez, Cyril Labbé. Controlled generation of synthetic corpora for NLP evaluation. 1st Workshop on Data-to-text Generation, Mar 2015, Edinburgh, United Kingdom. hal-03177929

HAL Id: hal-03177929
https://hal.science/hal-03177929

Submitted on 23 Mar 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Controlled generation of synthetic corpora for NLP evaluation

Jérémie Démarchez , Cyril Labbé
Université Grenoble Alpes - LIG
nom.prenom@e.ujf-grenoble.fr

Abstract

Automatic processing is mandatory to build a global and fair view of opinions and sentiments expressed on the web through comments and reviews. Various Extracting Tools (ETs) exists to automatically analyse comments and reviews; however checking the accuracy of such tools remain quite challenging. We propose a new approach for that purpose. The main idea is to use a data-to-text approach to generate a synthetic corpus which can be used to validate ETs. The data represent *what has to be said in which proportion about something* (i.e: 45% of the review says *the room is small*). A set of reviews (the synthetic corpus) is then generated and the correctness of an ET can then be assessed in regards to its fairness regarding the original data.

1 Introduction

In recent years, the amount of comments left on the internet exploded. These textual data are potentially very rich sources of information to find out the opinions of users. However, given the large amount of comments that describe an item, it appears complex and tedious to browse all of them to get a general idea. That is why, Extracting Tools (ETs) (Salvetti et al., 2004; Jing et al., 1998; Popescu and Etzioni, 2005; Labbé and Portet, 2012; Rahayu et al., 2010) have been built to generate summaries of the opinions expressed through the comments. Checking the accuracy and soundness of such approches remain difficult. The current way is to use corpora of comments that have been previously tagged by experts. These corpora are used as input for the ETs so to check and measure the accuracy of the extracted data. Putting aside the fact that experts often disagree, the annotation task by an expert panel is expensive and concerns a relatively small set of texts.

Thus, it appears useful to develop a method to check the relevance of data produced by ETs. The approach proposed here take advantage of the generation of a synthetic corpus. In this corpus *what is said at which frequency* is controlled so that the ETs can be tested on large and various corpora. A data-to-text approach is used to build a *controlled synthetic corpus*. This approach can be useful in many other natural language processing tasks (POS, translation, entity extraction, ...).

2 Method

A very simple way of generating the sought corpora would be to use a hand written probabilistic context free probabilistic grammar. Unfortunately, this kind of grammar would generate very repetitive reviews unless putting the burden on the writer of the grammar. That is why we propose an approach that *learn* the variety of existing natural corpora, by *learning* two kind of model language: 1) a probabilistic context-free grammar (F. Jelinek, 1992; Chomsky, 1956) and 2) a bi-gram model (Barbieri et al., 2012).

Learning sentences structures and word bi-gram. Prior to the generation of the synthetic corpus a probabilistic context-free grammar is learn on a natural corpus so to generate various kind of sentences. During this step, the resources necessary for text generation are created. Relevant resources are the probabilistic context-free grammar (Klein and Manning, 2003) and the statistics on bi-gram (a markov chains). The statistics regarding bi-gram are collected according the couples (word, tag).

Input data file As input of the generation process a *data file* is given. This file contains the number of comments/reviews to generate, a list of features to describe and for each feature, a list of

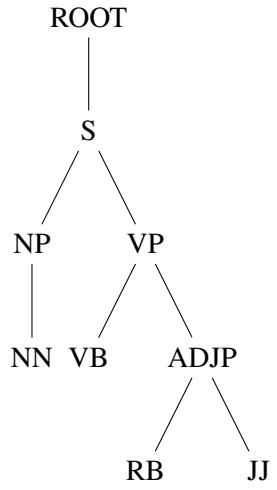

Figure 1: A synthetic sentence structure generated thanks to a learned PCFG.

adjectives describing it. This is given as a set of couples *(feature, probability)* and for each feature a set of *(adjective, probability)*. Which are respectively the probability of appearance of a feature and the probability of using a particular adjective in references to a feature.

Generating synthetic sentence structure The learned PCFG is then used to generate the synthetic sentence structure. Random trees like the one presented in figure 1 are obtained.

Choosing what to say at which frequency. First the subject of a sentence is randomly chosen and then inserted in the tree. In the same way, the adjective to use in the sentence is chosen. Then, we will use Markov chains on couple (words, tags) to choose the other words of the sentence among the ones from those revealed during learning.

Surface realization. The words are inserted under a lemmatized form in order to allow the realization of the sentence using a surface realizer (Gatt and Reiter, 2009) to obtain a proper sentence.

Quality of the generated texts An example of the process to generated a sentence is given in example 2.1. The first set of experiment shows that the quality of the generated texts mainly depends on the training corpus. This can be explained by the fact that when generated sentence structures are too complex the quality of produced sentences drops. The transformation of the tree generated using the PCFG must also be transform into a tree that fits the surface realizer abilities.

Example 2.1. 1. A morpho-syntactic tree is generated

```
(ROOT (S (NP (DT word1)(NN
word2))(VP (VBD word3)(ADJP
(JJ word4)))(. .)))
```

2. Random selection of a feature (i.e. pool) and an adjective (i.e. wide).

3. In the morpho-syntactic tree, the subject of the sentence is set in conjunction with the adjective to be used.

```
(ROOT (S (NP (DT word1) (NN
pool)) (VP (VBD word3) (ADJP
(JJ large))) (. .)))
```

4. All other words in the sentence are chosen using Markov chains.

```
(ROOT (S (NP (DT the) (NN
pool)) (VP (VBD be) (ADJP (JJ
large))) (. .)))
```

5. The realization is achieved (using SimpleNLG (Gatt and Reiter, 2009)), to obtain a grammatically correct English sentence. The pool was large.

3 Conclusion

The generated texts generally expressed the good informations, but not systematically. Similarly, they are generally grammatically correct, but there are still errors due to the parser used and the method of realization with SimpleNLG which is not suitable for some cases. In addition, when using a restrictive grammar, produced sentence structure does not vary enough in comparison to what is found in real corpora of comments. However, if we use a more permissive grammar, the risk of not controlling the expressed information is increased. Eventually, regarding the global implementation method, we can consider using it, if we have a adapted learning corpus, to evaluate certain automatic analysis tools. Indeed, the products comments are composed of simple sentences expressing clear information. So, if an automatic analysis tool to assess cannot produced a relevant summary, its effectiveness in front of a real corpus of comments could be doubtful.

References

Gabriele Barbieri, François Pachet, Pierre Roy, and Mirko Degli Esposti. 2012. Markov constraints for generating lyrics with style. In *ECAI*, volume 242, pages 115–120.

Noam Chomsky. 1956. Three models for the description of language. *IEEE Transactions on Information Theory*, 2(2):113–124.

R. L. Mercer F. Jelinek, J. D. Lafferty. 1992. Basic methods of probabilistic context free grammars. *Speech Recognition and Understanding NATO ASI Series*, 75:345–360.

Albert Gatt and Ehud Reiter. 2009. Simplenlg: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, ENLG '09, pages 90–93, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hongyan Jing, Regina Barzilay, Kathleen McKeown, and Michael Elhadad. 1998. Summarization evaluation methods: Experiments and analysis. In *IN AAAI SYMPOSIUM ON INTELLIGENT SUMMARIZATION*, pages 60–68.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *41st Meeting of the Association for Computational Linguistics*, pages 423–430.

Cyril Labbé and Francois Portet. 2012. Towards an abstractive opinion summarisation of multiple reviews in the tourism domain. In *SDAD 2012, The 1st International Workshop on Sentiment Discovery from Affective Data*, pages 87–94, september.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 339–346, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dwi Rahayu, Shonali Krishnaswamy, Cyril Labbé, and Oshadi Alahakoon. 2010. Web services for analysing and summarising online opinions and reviews. In *ServiceWave*.

Franco Salvetti, Stephen Lewis, and Christoph Reichenbach. 2004. Automatic opinion polarity classification of movie reviews. *Colorado Research in Linguistics*, 17(1).