Deep Unsupervised Extraction of knowledge bases from a dialogue corpus

Supervisor: Lina M. Rojas Barahona PhD, researcher on Dialogue, NADIA team at Orange-Labs.

ABSTRACT

The corpus DATCHA is a large dialogue corpus of human-human conversations between operators and clients from web chat services offered by customer-care centres. As part of the ANR funded project DATCHA *, the corpus has been annotated at different linguistic levels: morphological, syntax, semantics, discourse[1]. Nevertheless, there is a lack of domain specific annotations describing in detail the situation under discussion. For instance the problem that needs to be solved, the set of questions that help the diagnostic (the check list followed by the operator) and the proposed solutions. Since domain specific ontologies (i.e. a set of concepts and categories in a domain that shows their properties and their relations) are central in automatic dialogue systems, the goal of the internship is to apply unsupervised deep learning for extracting domain specific knowledge. In particular, we can take advantage of the evidence provided by the annotations. For instance, dialogue-act annotations, which describe in which part of the dialogue (in which turn or sequence of turns) the problem description and plan proposal are discussed.

Background

Many methods for unsupervised learning applied to natural language processing (NLP) have been proposed by the literature, such as clustering, co-clustering[2], distributional semantics[3], latent semantic analysis (LSA)[4] and Latent Dirichlet Allocation (LDA)[5].

Recently, deep learning emerged with unsupervised solutions, for instance, the famous word embeddings, word2vec[6] for semantic similarity. Other common unsupervised techniques used in deep learning are: Restricted Boltzmann Machines[7] and Variational Autoencoders[8].

The goal of the internship is to implement an unsupervised solution using deep learning to extract domain specific concepts (e.g. describing the services, the diagnostic check list and the proposed solutions). The obtained results must be compared to in house solutions namely co-clustering[2] and other publicity available solutions such as LDA.

References

1. Damnati, G., Guerraz, A. & Charlet, D. Web chat conversations from contact centers: a descriptive study. In *LREC* (2016).

2. Boullé, M. Data grid models for preparation and modeling in supervised learning. In Guyon, I., Cawley, G., Dror, G. & Saffari, A. (eds.) *Hands-On Pattern Recognition: Challenges in Machine Learning, volume 1*, 99–130 (Microtome Publishing, 2011).

3. Lund, K. & Burgess, C. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behav. research methods, instruments, & computers* **28**, 203–208 (1996).

4. Dumais, S. T. Latent semantic analysis. *Annu. review information science technology* **38**, 188–230 (2004).

5. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *J. machine Learn. research* **3**, 993–1022 (2003).

6. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

7. Larochelle, H. & Bengio, Y. Classification using discriminative restricted boltzmann machines. In *Proceedings of the 25th international conference on Machine learning*, 536–543 (ACM, 2008).

8. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[0] http://datcha.lif.univ-mrs.fr/