Text Normalization and Enrichment

Proposta de Dissertação de Mestrado

Orientador: Sebastião Pais
Co-Orientador: João Paulo Cordeiro

Departamento de Informática @ UBI
2019/2020

1 Context

People tweet more than 500 Million times daily, yielding a noisy, informal, but sometimes informative corpus of 140-character messages that mirrors the zeitgeist in an unprecedented manner. In the domain of surveillance systems, one is interested in extracting the specific information about “who" is doing “what", “where" and “when", that is conveyed in these short messages. This task is commonly called named entity recognition and aims to automatically extract mentions of rigid designators from text belonging to named-entity types such as persons, organizations, events, locations and timexes.

2 Goals

In this thesis, we are concerned with the improvement and robustness of named entity recognition systems for Twitter starting by following a line like the recent work of Saha *et al.* (2015) with biomedical texts. Indeed, most existing systems learn classifiers based on a small amount of labeled data. Although strong results can be obtained through cross-validation, these systems may not scale up in real-world environments such as social media due to high variety of unseen text contents. One solution to this problem is to apply ensemble learning techniques to take advantage of different learning paradigms such as conditional random fields or support vector machines Joachims (1998) that may combine them in some optimum consensus. By doing so, we expect that robust classifiers can be built and used reliably in a real-word environment such as Twitter.

To bridge the gap between unstructured text and structured machine readable knowledge bases, entity linking is performed and consists in mapping each entity mention in a tweet to a unique entity, i.e. an entry ID of a knowledge base such as Wikipedia or YAGO Suchanek *et al.* (2007). As such, each tweet is not an isolated segment of text but instead links to a knowledge base, which allows multilingual reasoning. A great deal of studies has been tackling entity linking Ferragina et Scaiella (2010) and more recently entity linking for social media texts Liu *et al.* (2013). Tweets pose special challenges to entity linking. First, a tweet is often too concise and too noisy to provide enough information

for similarity computing. Second, tweets have rich variations of named entities, and many of them fall out of the scope of the knowledge bases. Within the scope of this work, we propose to tune the strategy used for robust entity linking in Hoffart *et al.* (2011) for social media texts. We study the introduction of named entity continuous space Lin *et al.* (2015) in the disambiguation process. A promising new possibility can be the introduction of the recent work of Brazdil *et al.* (2015) in the domain of affinity mining, specially to discover and resolve apparently unrelated entities, mentioned in the social networks.

3 Tasks

1. Review the State-of-the-art, write the survey about this problematic;
2. Propose a new Unsupervised and Language Independent Approach to Entity Recognition;
3. Implementation;
4. Testing and evaluation;
5. The writing of the dissertation.

4 Contactos

Sebastião Pais (sebastiao@di.ubi.pt) - Gabinete 4.1
João Paulo Cordeiro (jpaulo@di.ubi.pt) - Gabinete 4.3

UBI, Departamento de Informática
Rua Marquês d'Ávila e Bolama
6201-001 Covilhã

Bibliography

Brazdil, P., Trigo, L., Cordeiro, J., Sarmento, R. et Valizadeh, M. (2015). Affinity mining of documents sets via network analysis, keywords and summaries. *Oslo Studies in Language*, 7(1).

Ferragina, P. et Scaiella, U. (2010). Tagme: On-the-fly annotation of short text fragments (by wikipedia entities). *In Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1625–1628, New York, NY, USA. ACM.

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S. et Weikum, G. (2011). Robust disambiguation of named entities in text. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 782–792, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *In Proceedings of the 10th European Conference on Machine Learning*, ECML'98, pages 137–142, Berlin, Heidelberg. Springer-Verlag.

Lin, Y., Liu, Z., Sun, M., Liu, Y. et Zhu, X. (2015). Learning entity and relation embeddings for knowledge graph completion. *In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2181–2187. AAAI Press.

Liu, X., Li, Y., Wu, H., Zhou, M., Wei, F. et Lu, Y. (2013). Entity linking for tweets. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1304–1311. Association for Computational Linguistics.

Saha, S., Ekbal, A. et Sikdar, U. K. (2015). Named entity recognition and classification in biomedical text using classifier ensemble. *Int. J. Data Min. Bioinformatics*, 11(4):365–391.

Suchanek, F. M., Kasneci, G. et Weikum, G. (2007). Yago: A core of semantic knowledge. *In Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA. ACM.