

IRISA participation to BioNLP-ST 2013: lazy-learning and information retrieval for information extraction tasks

Vincent Claveau

IRISA – CNRS

Campus de Beaulieu, 35042 Rennes, France

vincent.claveau@irisa.fr

Abstract

This paper describes the information extraction techniques developed in the framework of the participation of IRISA-TeXMex to the following BioNLP-ST13 tasks: Bacterial Biotope subtasks 1 and 2, and Graph Regulation Network. The approaches developed are general-purpose ones and do not rely on specialized pre-processing, nor specialized external data, and they are expected to work independently of the domain of the texts processed. They are classically based on machine learning techniques, but we put the emphasis on the use of similarity measures inherited from the information retrieval domain (Okapi-BM25 (Robertson et al., 1998), language modeling (Hiemstra, 1998)). Through the good results obtained for these tasks, we show that these simple settings are competitive provided that the representation and similarity chosen are well suited for the task.

1 Introduction

This paper describes the information extraction techniques developed in the framework of the participation of IRISA-TeXMex to BioNLP-ST13. For this first participation, we submitted runs for three tasks, concerning entity detection and categorization (Bacterial Biotope subtask 1, BB1), and relation detection and categorization (Bacterial Biotope subtask 2, BB2, and Graph Regulation Network, GRN).

Our participation to the BioNLP shared tasks takes place in the broader context of our work in the Quaero research program¹ in which we aim at developing fine grained indexing tools for

¹See www.quaero.org for a complete overview of this large research project.

multimedia content. Text-mining and information extraction problems are thus important issues to reach this goal. In this context, the approaches that we develop are general-purpose ones, that is, they are not designed for a specific domain such as Biology, Medecine, Genetics or Proteomics. Therefore, the approaches presented in this paper do not rely on specialized pre-processing, nor specialized external data, and they are expected to work independently of the domain of the texts processed.

The remaining of this paper is structured as follows: the next section presents general insights on the methodology used throughout our participation, whatever the task. Sections 3, 4 and 5 respectively describe the techniques developed and their results for BB1, BB2 and GRN. Last, some conclusive remarks and perspectives are given in Section 6.

2 Methodological corpus

From a methodological point of view, our approaches used for these tasks are machine learning ones. Indeed, since the first approaches of information extraction based on the definition of extraction patterns (Riloff, 1996; Soderland, 1999), using surface clues or syntactic and semantic information (Miller et al., 2000), machine learning techniques have shown high performance and versatility. Generally, the task is seen as a supervised classification one: the training data are used to infer a classifier able to handle new, unlabeled data. Most of the state-of-the-art techniques adopt this framework, but differ in the kind of information used and on the way to use it. For instance, concerning the syntactic information, different representations were studied: sequences or sub-sequences (Culotta et al., 2006; Bunescu and Mooney, 2006), shallow parsing (Pustejovsky et al., 2002; Zelenko et al., 2003), dependencies (Manine et al., 2009), trees (Zhang et al., 2006; Liu et al., 2007), graphs (Culotta and Sorensen,

2004; Fundel et al., 2007), etc. Also, exploiting semantic information, for instance through distributional analysis seems promising (Sun et al., 2011).

The approaches also differ in the inference techniques used. Many were explored, like neural networks (Barnickel et al., 2009) or logistic regression (Mintz et al., 2009), but those relying on a metric space search, such as Support Vector Machines (SVM) or k-Nearest Neighbours (kNN) are known to achieve state-of-the-art results (Zelenko et al., 2003; Culotta and Sorensen, 2004). The crux of the matter for these methods is to devise a good metric between the objects, that is, a good kernel. For instance, string kernels (Lodhi et al., 2002) or graph kernels (Tikk et al., 2012) have shown interesting performance.

Our approaches in these shared tasks also adopt this general framework. In particular, they are chiefly based on simple machine learning techniques, such as kNN. In this classification technique, new instances whose classes are unknown are compared with training ones (instances with known classes). Among the latter, the closest ones in the feature space are used to decide the class of the new instance, usually by a majority vote. Beyond the apparent simplicity of this machine learning technique, the heart of the problem relies in the two following points:

- using a relevant distance or similarity measure in the feature space to compare the instances;
- finding the best voting process (number of nearest neighbors, voting modalities...)

There is no real training step *per se*, but kNN is truly a machine learning approach since the inductive step is made when computing the similarity and the vote for the classification of a new instance, hence the qualification of 'lazy-learning' method.

In our work, we explore the use of similarity measures inherited from the information retrieval (IR) domain. Indeed, IR has a long history when it comes to comparing textual elements (Rao et al., 2011) which may offer new similarity measures for information extraction either for kernel-based methods or, in our case, for kNN. Therefore, in the remaining of the article, we mainly describe the choice of this similarity measure, and adopt the standard notation used in IR to denote a

similarity function: RSV (Retrieval Status Value, higher score denotes higher similarity). In practice, all the algorithms and tools were developed in Python, using NLTK (Loper and Bird, 2002) for basic pre-processing.

3 Term extraction and categorization: Bacteria Biotope sub-task 1

This section describes our participation to sub-task 1 of the Bacteria Biotope track. The first sub-section presents the task as we interpreted it, which explains some conceptual choices of our approach. The latter is then detailed (sub-section 3.2) and its results are reported (sub-section 3.3).

3.1 Task interpretation

This task aims at detecting and categorizing entities based on an ontology. This task has some important characteristics:

- it has an important number of categories;
- categories are hierarchically organized;
- few examples for each categories are given through the ontology and the examples.

Moreover, some facts are observed in the training data:

- entities are mostly noun phrase;
- most of the entities appear in a form very close to their corresponding ontology entry.

Based on all these considerations and to our point of view explained in the previous section, this task is interpreted as an automatic categorization one: a candidate (portion of the analyzed text) is assigned an ontological category or a negative class (stating) that the candidate does not belong to any spotted category.

In the state-of-the-art, such problems are often considered as labeling ones for which stochastic techniques like HMM, MaxEnt models, or more recently CRF (Lafferty et al., 2001), have shown very good results in a large variety of tasks (Wang et al., 2006; Pranjali et al., 2006, *inter alia*). Yet, for this specific case, these techniques do not seem to be fully suited for different reasons:

- a very high number of possible classes is to be handled, which may cause complexity problems;
- the size of the training set is relatively small compared to the size of the label set;

- embedding external knowledge (i.e. the ontology), for instance as features, cannot be done easily.

On the contrary of these stochastic methods, our approach does not rely on the sequential aspect of the problem. It is based on lazy machine learning (kNN), detailed hereafter, with a description allowing us to make the most of the ontology and the annotated texts as training data.

3.2 Approach

In the approach developed for this task, the context of a candidate is not taken into account. We only rely on the internal components (word-forms) of the candidate to decide whether it is an entity and what is its category. That is why both the ontology and the given annotated texts are equally considered as training data.

More precisely, this approach is implemented in two steps. In the first step, the texts are searched for almost exact occurrences of an ontology entry. Slight variations are allowed, such as case, word insertion and singular/plural forms. In practice, this approach is implemented with simple regular expressions automatically constructed from the ontology and the annotated texts.

In the second step, a more complex processing is undergone in order to retrieve more entities (to improve the recall rate). It relies on a 1-nearest neighbor classification of the noun phrase (NP) extracted from the text. A NP chunker is built by training a MaxEnt model from the CONLL 2000 shared task dataset (articles from the Wall Street Journal corpus). This NP chunker is first applied on the training data. All NP collected that do not belong to any wanted ontological categories are kept as examples of a negative class. The NP chunker is then applied to the test data. Each extracted NP is considered as a candidate which is compared with the ontological entries and the collected negative noun phrases. This candidate finally receives the same class than the closest NP (i.e. the ontological category identifier or the negative class).

As explained in the previous section, the keystone of such an approach is to devise an efficient similarity measure. In order to retrieve the closest known NP, we examine the word-forms composing the candidate, considered as a bag-of-words. An analogy is thus made with information retrieval: ontological categories are considered as documents, and the candidate is considered as a

query. A similarity measure inherited from information retrieval, called Okapi-BM25 (Robertson et al., 1998), is used. It can be seen as a modern variant of TF-IDF/cosine similarity, as detailed in Eqn. 1 where t is a term occurring qtf times in the candidate q , c a category (in which the term t occurs tf times), $k_1 = 2$, $k_3 = 1000$ and $b = 0.75$ are constants, df is the document frequency (number of categories in which t appears), dl is the document length, that is, in our case the number of words of the terms in that category, dl_{avg} is the average length (number of words) of a category.

$$RSV(q, c) = \sum_{t \in q} qTF(t) * TF(t, c) * IDF(t) \quad (1)$$

with:

$$qTF(t) = \frac{(k_3 + 1) * qtf}{k_3 + qtf}$$

$$TF(t, c) = \frac{tf * (k_1 + 1)}{tf + k_1 * (1 - b + b * dl(c)/dl_{avg})}$$

$$IDF(t) = \log \frac{N - df(t) + 0.5}{df(t) + 0.5}$$

Finally, the category c^* for the candidate q is chosen among the set \mathcal{C} of all the possible ones (including the negative category), such that:

$$c^* = \arg \max_{c \in \mathcal{C}} RSV(q, c)$$

The whole approach is illustrated in Fig. 1.

Still in order to improve recall, unknown words (words that do not appear in any category) undergo an additional process. The definition of the word in WordNet, if present, is used to extend the candidate, in a very similar way to what would be query expansion (Voorhees, 1998). In case of polysemous words, the first definition is used.

3.3 Results

Figure 2 presents the official results of the participating teams on the test dataset. Our approach obtains good overall performance compared with other team's results and ranks first in terms of Slot Error Rate (SER, combining the number of substitution S, insertion I, deletion D and Matches M). As it appears, this is mainly due to a better recall rate. Of course, this improved recall has its drawback: the precision of our approach is a bit lower than some of the other teams. This is confirmed

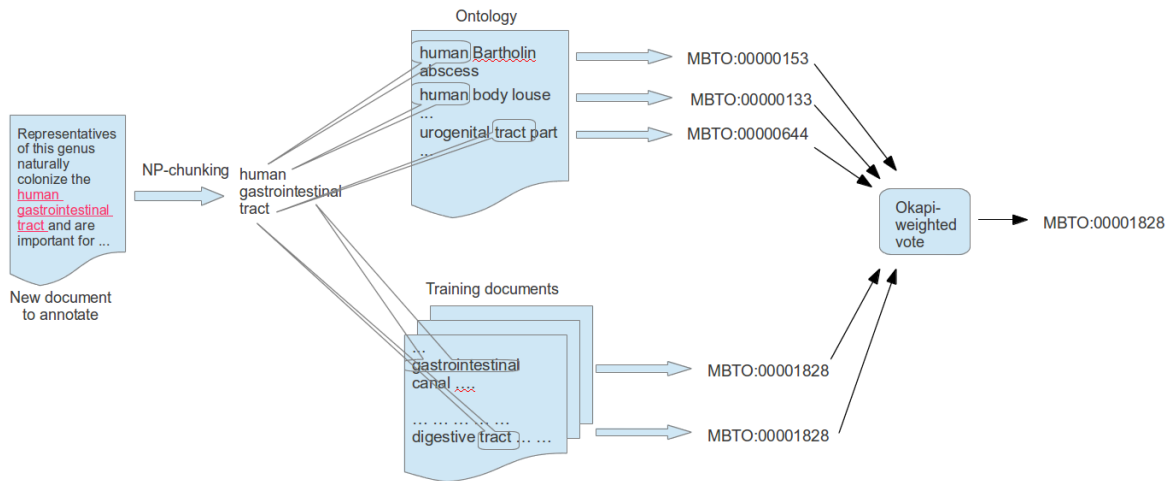


Figure 1: k-NN based approach based on IR similarity measures

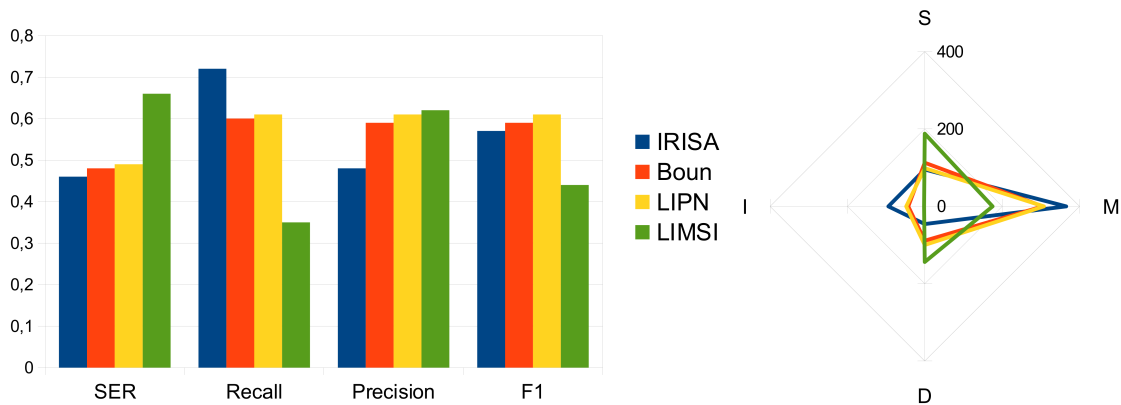


Figure 2: BB1 official results: global performance rates (left); error analysis (right)

by the general shape of our technique compared with others' (Figure 2, right) with more matches, but also more insertions.

In order to analyze the performance of each component, we also report results of step 1 (quasi-exact matches with regular expression) alone, step 2 alone, and a study of the influence of using WordNet to extend the candidate. The results of these different settings, on the development dataset, are given in Figure 3. From these results, the first point worth noting is the difference of overall performance between the development set and the test set (SER on the latter is almost two times higher than on the former). Yet, without access to the test set, a thorough analysis of this phenomenon cannot be undergone. Another striking point is the very good performance of step 1, that

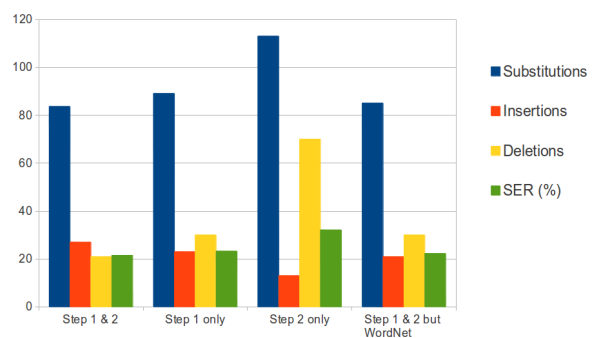


Figure 3: Influence of each extraction component

is, the simple search for quasi identical ontology phrases in the text. Compared to this, step 2 performs worse, with many false negatives (deletions) and misclassifications (substitutions). A close examination of the causes of these errors reveals that the IR-based classification process is not at fault, but it is misled by wrong candidates proposed by the NP chunker. Besides the problem of performance of our chunker, it also underlines the limit of our hypothesis of using only noun phrases as possible candidates. In spite of these problems, step 2 provides complementary predictions to step 1, as their combination obtains better results than each one. This is also the case with the WordNet-based expansion, which brings slightly better results.

4 Extracting relation: Bacteria Biotope sub-task 2

This section is dedicated to the presentation of our participation to Bacteria Biotope sub-task 2. As for the sub-task 1, we first present the task as we interpreted it, then the approach, and last some results.

4.1 Task interpretation

This task aims at extracting and categorizing localization and part-of relations that may be reported in scientific abstracts between Bacteria, Habitat and Geographical spots. For this particular sub-task, the entities (boundaries in the text and type) were provided.

As explained in Section 2, expert approaches based on hand-coded patterns are outperformed by state-of-the-art studies which consider this kind of tasks as a classification one. Training data help to infer a classifier able to decide, based on features extracted from the text, whether two entities share a relation, and able to label this relation if needed. We also adopt this framework and exploit a system developed in-house (Ebadat, 2011) which has shown very good performance on the protein-protein-interaction task of the LLL dataset (Nédellec, 2005). From a computational point of view, two directed relations are to be considered for this task, plus the 'negative' relation stating that no localization or part-of relation exists between the entities. Therefore, the classifier has to handle five labels.

4.2 Approach

The extraction method used for this task only exploits shallow linguistic information, which is easy to obtain and ensures the necessary robustness, while providing good results on previous tasks (Ebadat, 2011). One of its main interests is to take into account the sequential aspect of the task with the help of n-gram language models. Thus, a relation is represented by the sequence of lemmas occurring between the agent and the target, if the agent occurs before the target, or between the target and the agent otherwise. A language model is built for each example Ex , that is, the probabilities based on the occurrences of n-grams in Ex are computed; this language model is written \mathcal{M}_{Ex} . The class (including the 'negative' class) and direction (left-to-right, LTR or right-to-left, RTL) of each example is also memorized.

Given a relation candidate (that is, two proteins or genes in a sentence), it is possible to evaluate its proximity with any example, or more precisely the probability that this example has generated the candidate. Let us note $C = \langle w_1, w_2, \dots, w_m \rangle$ the sequence of lemmas between the proteins. For n-grams of n lemmas, this probability is classically computed as:

$$P(C|\mathcal{M}_{Ex}) = \prod_{i=1}^m P(w_i|w_{i-n}..w_{i-1}, \mathcal{M}_{Ex})$$

As for any language model in practice, probabilities are smoothed in order to prevent unseen n-grams to yield 0 for the whole sequence. In the experiments reported below, we consider bigrams of lemmas. Different strategies for smoothing are used: as it is done in language modeling for IR (Hiemstra, 1998), probabilities estimated from the example are linearly combined with those computed on the whole set of example for this class. In case of unknown n-grams, an interpolation with lower order n-grams (unigram in this case) combined with an absolute discounting (Ney et al., 1994) is performed.

In order to prevent examples with long sequences to be favored, the probability of generating the example from the candidate ($P(Ex|\mathcal{M}_C)$) is also taken into account. Finally, the similarity between an example and a candidate is:

$$RSV(Ex, C) = \min(P(Ex|\mathcal{M}_C), P(C|\mathcal{M}_{Ex}))$$

The class is finally attributed to the candidate by a k-nearest neighbor algorithm: the k most sim-

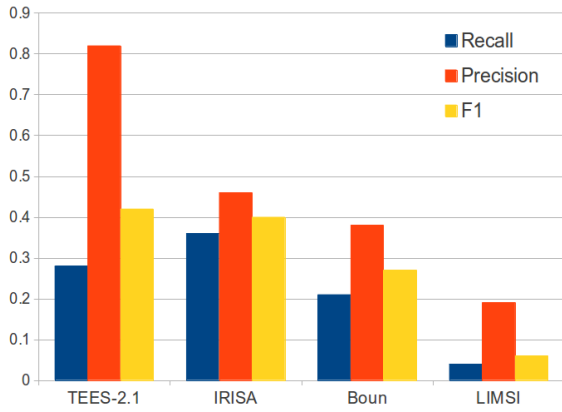


Figure 4: BB2 official results in terms of recall, precision and F-score

ilar examples (highest RSV) are calculated and a majority vote is performed. For this task, k was set to 10 according to cross-validation experiments. This lazy-learning technique is expected to be more suited to this kind of tasks than the model-based ones (such as SVM) proposed in the literature since it better takes into account the variety of ways to express a relation.

4.3 Results

The official results are presented in Figure 4. In terms of F-score, our team ranks close second, but with a different recall/precision compromise than TEES-2.1. The detailed results provided by the organizers show that no Part-of relations are retrieved. From the analysis of errors on the development set, it appears that the simplicity of our representation is at fault in most cases of misclassifications. Indeed, important keywords frequently occur outside of the sub-sequence delimited by the two entities. The use of syntactic information, as proposed for the GRN task in the next section, is expected to help overcome this problem.

5 Extracting relation: regulation network

5.1 Task interpretation and approach

Despite the different application context and the different evaluation framework, we consider this relation extraction task in a similar way than in the previous section. Therefore, we use the same approach already described in Section 4.2. Yet, instead of using the sequence of lemmas between the entities, we rely on the sequence built from the

Activation of T1 requires expression of T2, coding for a protein of T3 family, and of the mother cell-specific T4.

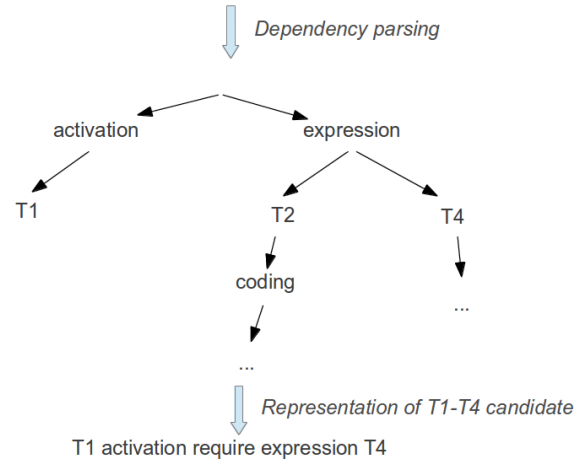


Figure 5: Example of syntactic representation used for the GRN task

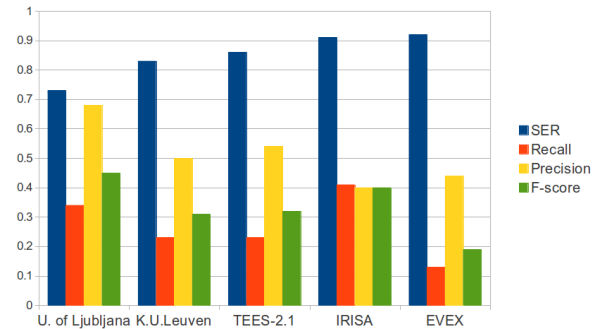


Figure 6: GRN official results in terms of strict Slot Error Rate (SER), recall, precision and F-score

shortest syntactic path between the entities as it is done in many studies (Manine et al., 2009, *inter alia*). The text is thus parsed with MALT parser (Nivre, 2008) and its pre-trained Penn Treebank model (Marcus et al., 1993). The lemmas occurring along the syntactic path between the entities, from the source to the target, are collected as illustrated in Figure 5.

5.2 Results

The official results reported in Fig. 6 shows that although our approach only ranks fourth in terms of Slot Error Rate (SER), its general performance is competitive in terms of Recall and F-score, but its relatively lower precision impacts the global SER score. It is also interesting to consider a relaxed version of these evaluation measures in which sub-

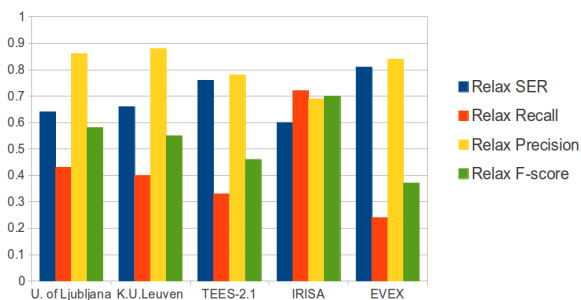


Figure 7: GRN official results in terms of relaxed Slot Error Rate (SER), recall, precision and F-score

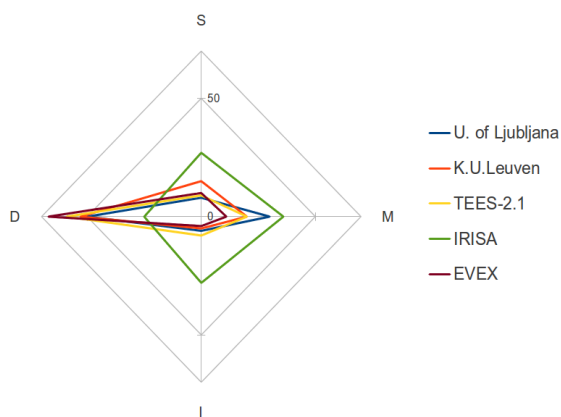


Figure 8: Analysis of errors of the GRN task

stitutions are not penalized. It therefore evaluates the ability of the methods to build the regulation network whatever the real relation between entities. As it appears in Figure 7, in that case, our approach brings the best results in terms of F-score and SER. As for the BB2 task, it means that the pro-eminent errors are between labels of valid relations, but not on the validity of the relation. This is also noticeable in Figure 8 in which the global profile of our approach underlines its capacity to retrieve more relations, but also to generate more substitution and insertion errors than the other approaches. The complete causes of these misclassifications are still to be investigated, but a close examination of the results shows two possible causes:

- the parser makes many mistakes on conjunction and prepositional attachment, which is especially harmful for the long sentences used in the dataset;
- our representation omits to include negation or important adverbs, which by definition are

not part of the shortest path, but are essential to correctly characterize the relation.

The first cause is not specific to these data and is a well-known problem of parsing, but hard to overcome at our level. The second cause is specific to our approach, and militate, to some extents, to devise a more complex representation than the shortest path one.

6 Conclusion and future work

For this first participation of IRISA to BioNLP shared tasks, simple models were implemented, using no domain-specific knowledge. According to the task, these models obtained more or less good rankings, but all have been shown to be competitive with other teams' results. Our approaches put the emphasis on the similarity computing between known instances instead of complex machine learning techniques. By making analogies with information retrieval, this similarity aims at being the most relevant for the considered task and at finding the closest known examples of any new instance to be classified.

For instance, we made the most of the vector-space measure Okapi-BM25 combined with a bag-of-word representation for the first sub-task of Bacterial Biotope, and of the language modeling adapted from (Hiemstra, 1998) for the sequential representation used in the second sub-task of Bacterial Biotope and for Gene Regulation Network.

Many parameters as well as other similarity choices have not been explored due to the short delay imposed by the challenge schedule. As a future work, it would be interesting to automatically set these parameters according to the data. In particular, a complex version of the BM-25 RSV function permits to include relevance feedback, which, in our machine learning framework, corresponds to using training data to adapt the BM-25 formula. Another research avenue concerns the synonymy/paraphrasing problem, which is not correctly handled by our word-based methods. Thus, semantic analysis techniques used in IR (and other NLP domains) such as Latent Semantic Indexing (Deerwester et al., 1990) or Latent Dirichlet Allocation (Blei et al., 2003) may also lead to interesting results.

Acknowledgment

This work was partly funded by OSEO, the French agency for innovation, in the framework of the

Quaero project.

References

- [Barnickel et al.2009] T. Barnickel, J. Weston, R. Collobert, H.W. Mewes, and V. Stümpflen. 2009. Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts. *PLoS One*, 4(7).
- [Blei et al.2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March.
- [Bunescu and Mooney2006] R. Bunescu and R. Mooney. 2006. Subsequence kernels for relation extraction. *Advances in Neural Information Processing Systems*, 18.
- [Culotta and Sorensen2004] A. Culotta and J. Sorensen. 2004. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*.
- [Culotta et al.2006] A. Culotta, A. McCallum, and J. Betz. 2006. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 296–303.
- [Deerwester et al.1990] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.
- [Ebadat2011] Ali Reza Ebadat. 2011. Extracting protein-protein interactions with language modelling. In *Proceeding of the RANLP Student Research Workshop*, pages 60–66, Hissar, Bulgaria.
- [Fundel et al.2007] K. Fundel, R. Küffner, and R. Zimmer. 2007. Relex – relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371.
- [Hiemstra1998] D. Hiemstra. 1998. A linguistically motivated probabilistic model of information retrieval. In *Proc. of European Conference on Digital Libraries, ECDL, Heraklion, Greece*.
- [Lafferty et al.2001] J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*.
- [Liu et al.2007] Y. Liu, Z. Shi, and A. Sarkar. 2007. Exploiting rich syntactic information for relation extraction from biomedical articles. In *Human Language Technologies 2007: Conf. North American Chapter of the Association for Computational Linguistics; Companion Volume (NAACL-Short’07)*, pages 97–100.
- [Lodhi et al.2002] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. 2002. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444.
- [Loper and Bird2002] Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1, ETMTNLP ’02*, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Manine et al.2009] A.-P. Manine, E. Alphonse, and P. Bessières. 2009. Learning ontological rules to extract multiple relations of genic interactions from text. *Int. Journal of Medical Informatics*, 78(12):31–38.
- [Marcus et al.1993] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19(2):313–330, June.
- [Miller et al.2000] S. Miller, H. Fox, L. Ramswhaw, and R. Weischedel. 2000. A novel use of statistical parsing to extract information from text. In *Proc. 1st North American Chapter of the Association for Computational Linguistics Conf.*, pages 226–233.
- [Mintz et al.2009] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proc. Joint Conf. 47th Annual Meeting of the ACL and the 4th Int. Joint Conf. on Natural Language Processing of the AFNLP*.
- [Ney et al.1994] Hermann Ney, Ute Essen, and Reinhard Kneser. 1994. On structuring probabilistic dependencies in stochastic language modelling. *Computer Speech and Language*, 8:1–38.
- [Nivre2008] Joakim Nivre. 2008. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34(4):513–553.
- [Nédellec2005] Claire Nédellec, editor. 2005. *Learning language in logic – Genic interaction extraction challenge*, in Proc. of the 4th Learning Language in Logic Workshop (LLL’05), Bonn, Germany.
- [Pranjal et al.2006] Awasthi Pranjal, Rao Delip, and Ravindran Balaraman. 2006. Part of speech tagging and chunking with hmm and crf. In *Proceedings of NLP Association of India (NLP AI) Machine Learning Contest*.
- [Pustejovsky et al.2002] J. Pustejovsky, J. Castano, J. Zhang, M. Kotecki, and B. Cochran. 2002. Robust relational parsing over biomedical literature: Extracting inhibit relations. In *Proceedings of the Pacific Symposium in Biocomputing*, pages 362–373.

- [Rao et al.2011] Delip Rao, Paul McNamee, and Mark Dredze. 2011. Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, Multilingual Information Extraction and Summarization*.
- [Riloff1996] E. Riloff. 1996. Automatically generating extraction patterns from untagged text. In *Proc. 13th Natl. Conf. on Artificial Intelligence (AAAI-96)*, pages 1044–1049.
- [Robertson et al.1998] Stephen E. Robertson, Steve Walker, and Micheline Hancock-Beaulieu. 1998. Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive. In *Proceedings of the 7th Text Retrieval Conference, TREC-7*, pages 199–210.
- [Soderland1999] S. Soderland. 1999. Learning information extraction rules for semi-structured and free text. *Machine Learning Journal*, 34(1-3):233–272.
- [Sun et al.2011] A. Sun, R. Grishman, and S. Sekine. 2011. Semi-supervised relation extraction with large-scale word clustering. In *Proc. 49th Annual Meeting of the Association for Computational Linguistics*, pages 521–529.
- [Tikk et al.2012] D. Tikk, P. Thomas, P. Palaga, J. Hakenberg, and U. Leser. 2012. A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. *PLoS Computing Biology*, 6(7).
- [Voorhees1998] E. Voorhees, 1998. *C. Fellbaum (ed.), WORDNET: An Electronic Lexical Database*, chapter Using WORDNET for Text Retrieval, pages 285–303. The MIT Press.
- [Wang et al.2006] Tao Wang, Jianguo Li, Qian Diao, Yimin Zhang, Wei Hu, and Carole Dulong. 2006. Semantic event detection using conditional random fields. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW '06)*.
- [Zelenko et al.2003] D. Zelenko, C. Aone, and A. Richardella. 2003. Kernel methods for relation extraction. *Journal of Machine Learning Research*, 3:1083–1106.
- [Zhang et al.2006] M. Zhang, J. Zhang, and J. Su. 2006. Exploring syntactic features for relation extraction using a convolution tree kernel. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 288–295.