# Using shallow linguistic features for relation extraction in bio-medical texts

Ali Reza Ebadat[1] Vincent Claveau[2] Pascale Sébillot[3]
(1) INRIA-INSA, (2) IRISA-CNRS, (3) IRISA-INSA
Campus de Beaulieu, 35042 Rennes, France
ali_reza.ebadat@inria.fr, vincent.claveau@irisa.fr, pascale.sebillot@irisa.fr

**Résumé.** Dans cet article[1], nous proposons de modéliser la tâche d'extraction de relations à partir de corpus textuels comme un problème de classification. Nous montrons que, dans ce cadre, des représentations fondées sur des informations linguistiques de surface sont suffisantes pour que des algorithmes d'apprentissage artificiel standards les exploitant rivalisent avec les meilleurs systèmes d'extraction de relations reposant sur des connaissances issues d'analyses profondes (analyses syntaxiques ou sémantiques). Nous montrons également qu'en prenant davantage en compte les spécificités de la tâche d'extraction à réaliser et des données disponibles, il est possible d'obtenir des méthodes encore plus efficaces tout en exploitant ces informations simples. La technique originale à base d'apprentissage « paresseux » et de modèles de langue que nous évaluons en extraction d'interactions géniques sur les données du challenge LLL2005 dépasse les résultats de l'état de l'art.

**Abstract.** In this paper[2], we model the corpus-based relation extraction task as a classification problem. We show that, in this framework, standard machine learning systems exploiting representations simply based on shallow linguistic information can rival state-of-the-art systems that rely on deep linguistic analysis. Even more effective systems can be obtained, still using these easy and reliable pieces of information, if the specifics of the extraction task and the data are taken into account. Our original method combining lazy learning and language modeling out-performs the existing systems when evaluated on the LLL2005 protein-protein interaction extraction task data.

**Mots-clés :** Extraction de relations, classification, apprentissage paresseux, modèle de langue, analyse linguistique de surface.

**Keywords:** Relation extraction, classification, lazy learning, langage model, shallow linguistic analysis.

[1]Ces travaux ont été réalisés dans le cadre du programme QUAERO, financé par OSEO, agence française pour l'innovation.
[2]This work was achieved as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

# 1 Introduction

Since the nineties, a lot of research work has been dedicated to the problem of corpus-based knowledge acquisition, whether the aimed knowledge is terminology, special cases of vocabulary (e.g. named entities), lexical relations between words or more functional ones. This paper focuses on this last kind of acquisition, i.e., relation extraction, and more specifically on Protein-Protein Interaction (PPI) extraction from bio-medical texts. The goal of PPI extraction is to find pairs of proteins within sentences such that one protein is described as regulating, inhibiting, or binding the other. In functional genomics, these interactions, which are not available in structured databases but scattered in scientific papers, are central to determine the function of the genes.

In order to extract PPIs, the texts which contain the interactions have to be analyzed. Two kinds of linguistic analysis can be performed for this purpose: deep and shallow. Automatic deep analysis, which provides a syntactic or semantic parsing of each sentence, can be a useful source of information. However, tools for automatic deep analysis are available only for a limited number of natural languages, and produce imperfect results. Manual deep analysis, on the other hand, is time consuming and expensive. Another way to analyze texts is to rely only on a shallow linguistic analysis, taking into account the sole words, lemmas or parts of speech (POS) tags. Automatic tools for shallow analysis are available for many languages, and are (sufficiently) reliable.

In this paper, we advocate the use of shallow linguistic features for relation extraction tasks. First, we show that these easy and reliable pieces of information can be efficiently used as features in a machine learning (ML) framework, resulting in good PPI extraction systems, as effective as many systems relying on deep linguistic analysis. Furthering this idea, we propose a new simple yet original system, called LM-kNN and based on language modeling, that out-performs the state-of-the-art systems.

The paper is organized as follows. Section 2 reviews related work on PPI extraction from bio-medical texts. Section 3 specifies the problem, the methodology and describes the LM-kNN technique. Results when using classical ML algorithms are given in Section 4, together with a comparison with existing systems. The last section presents a conclusion and some future work.

# 2 Related work

In this literature review, focus is set on researches dedicated to relation extraction from bio-medical texts, especially those evaluated in a PPI framework. The systems proposed for this task can be organized into different groups, depending on the source of knowledge (deep vs. shallow linguistic information) and on the approach used (manual vs. ML).

For instance, RelEx (Fundel *et al.*, 2007) exploits manually built extraction rules handling deep and shallow linguistic information. This system yields good results; yet using such an hand-elaborated knowledge is a bottleneck requiring expertise for any new domain. Thus, many ML-based approaches were proposed to overcome this limitation. The ML techniques range from SVM with complex kernels (Airola *et al.*, 2008; Kim *et al.*, 2010) or CRF (Li *et al.*, 2007), to expressive techniques like inductive logic programming (Phuong *et al.*, 2003). Words surrounding a pair of proteins or some of their linguistic features can be considered as shallow linguistic features to train the systems (Bunescu & Mooney, 2006; Giuliano *et al.*, 2006; Sun *et al.*, 2007). However, most of the best performing techniques rely on deep linguistic analysis like syntactic parsing. Indeed, grammatical relations are assumed to be important for PPI extraction, especially when few training data compared to test data are available (Fayruzov *et al.*, 2009). Yet, the performance of extraction systems being sensitive to the accuracy of automatic parsers (Fayruzov *et al.*, 2008), shallow linguistic information still remains an option (Xiao *et al.*, 2005), though up-to-now less effective than deep one.

In this work, we defend the hypothesis that shallow linguistic information combined with standard ML approaches is sufficient to reach good results. Furthermore, we propose a system demonstrating that, when this simple information is cleverly used, it even out-performs state-of-the-art systems.

# 3 Approach

In this section, we present the different machine learning approaches, based on shallow linguistic features, that we experimented. The first subsection present how we model the PPI task as a ML problem—and in particular how relations are described—and the classification tools commonly used for similar tasks. In the next subsection, we propose a new relation extraction technique, based on language modeling, which is expected to be more efficient than the existing ones.

### 3.1 Modeling the relation extraction task as a machine learning problem

In PPI extraction, the goal is to predict if there is any interaction between two proteins. In such a case, the relation is directed, that is, one of the entities is the agent, the other one the target. For example, in the sentence reported in Figure 1 in which entities (proteins) are in bold, there is a relation between *GerE* and *cotD* for which *GerE* is the agent and *cotD* is the target.

> ***GerE*** *stimulates* ***cotD*** *transcription and inhibits* ***cotA*** *transcription in vitro by* ***sigma K*** *RNA polymersase, as expected from in vivo studies, and, unexpectedly, profoundly inhibits in vitro transcription of the gene (****sigK****) that encode* ***sigma K****.*

Figure 1: Sample sentence of protein-protein interaction

To handle this directed relation problem, we model it as a 3-class ML task. For each training sentence, each pair of entities is either tagged as *None* if the entity pair does not concern any interaction, *LTR* if the interaction is from the left to the right (in the sentence word order), and *RTL* if the interaction is from the right to the left. A relation is simply represented by the bag of lemmas occurring between the two entities. This bag-of-lemmas representation of the training examples is exploited by popularly used ML techniques. In the experiments presented in section 4, we report the results obtained with the libSVM SVM implementation (Chang & Lin, 2001) and random forests (Breiman, 2001) as implemented in (Hall *et al.*, 2009).

### 3.2 Nearest neighbors with language modeling

Besides these somewhat classical machine learning approaches, we propose a new technique to extract relations. As the previous ones, it still uses shallow linguistic information, which is easy to obtain and ensures the necessary robustness. One of the main differences concerns the representation of the examples: it takes into account the sequential aspect of the task with the help of n-gram language models. Thus, an example consists in the sequence of words appearing between two entities in a training sentence. 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 (LTR, RTL or none) of each example is also memorized.

Given a relation candidate (that is, two proteins or genes in a test sentence), it is possible to evaluate its proximity with any example, or more precisely the probability that this example has generated the candidate. Intuitively, this probability represents how likely the sequence of words of the candidate can be built from the sequence of words of the example. Let us note $C =< w_1, w_2, ..., w_m >$ the sequence of lemmas between the two proteins. For $n$-grams (in our case $n$ 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 and simply use interpolation with lower order n-grams (unigram in this case) combined with an absolute discounting (Ney *et al.*, 1994).

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:

$$sim(Ex, C) = \min\left(P(Ex|\mathcal{M}_C), P(C|\mathcal{M}_{Ex})\right) \ .$$

The class is finally attributed to the candidate by a k-nearest neighbor algorithm: the 10 most similar examples (highest $sim$) are calculated and a majority vote is performed. This lazy-learning technique is expected to be more suited to this kind of tasks than the model-based ones proposed in the previous sub-section since it better takes into account the variety of ways to express a relation.

## 4 Experiments

This section presents experiments with the different relation extraction systems described above. The data used and the evaluation metrics and methodologies are first detailed. Then results are provided and discussed.

## 4.1 LLL data

To evaluate the extraction systems, the data developed for the Learning Language in Logic 2005 (LLL05) shared task (Nédellec, 2005) were used (see (Pyysalo *et al.*, 2008) for a presentation of the LLL05 corpora). The goal of LLL05 was to extract protein/gene interactions in abstracts from the Medline bibliography database.

The provided training set is composed of sentences in which a total of 161 interactions between genes/proteins are identified. All pairs of proteins in a sentence with no interaction between their constituents were considered as negative examples.

The test set is composed of another set of sentences for which the ground-truth is kept unknown; and the results are computed by submitting the predictions to a web service. The original LLL challenge offered the possibility to train and test the systems only on interactions expressed without the help of co-references (mostly with pronouns designating a previously mentioned entity). The training and test data were also provided with or without manual syntactic annotations of the sentences (dependency analysis). Of course, in order to evaluate our systems in a realistic way, we used the data containing interactions expressed with or without co-references, and we did not consider the manual syntactic annotation.

## 4.2 Evaluation

We evaluate our LM-kNN approach and compare it to the more traditional machine learning techniques and state-of-the-art systems for PPI extraction. The evaluation metrics chosen in our experiments are those classically used in this domain: precision, recall and f-measure. Partially correct answers, like an interaction between two entities correctly detected but with the wrong interaction direction, are considered as wrong answers.

Table 1 reports the performance obtained by these techniques on the complete test set (including both interactions expressed with or without co-references). For comparison purposes, the results on this dataset reported by other studies are also included. Since many teams only considered the evaluation without co-references—which is supposed to correspond to an easier task—, we also report the results of our LM-kNN approach and other state-of-the-art systems in this context in Table 2. The first part of each table concerns systems using raw data (no manual annotation) and the second part contains results of other systems using the manual syntactic analysis also available in the LLL data. The best-performing systems for each part are highlighted in grey.

| **System** | **Precision** | **Recall** | **F-measure** |
|---|---|---|---|
| systems using raw data | | | |
| (Goadrich *et al.*, 2005) | 25.0 | 81.4 | 38.2 |
| Random Forest | 57.9 | 48.1 | 52.6 |
| libSVM linear kernel | 58.0 | 56.6 | 57.3 |
| LM-kNN | 70.9 | 79.5 | 75 |
| systems using manually annotated data | | | |
| (Katrenko *et al.*, 2005) | 51.8 | 16.8 | 25.4 |
| (Goadrich *et al.*, 2005) | 14.0 | 93.1 | 24.4 |

Table 1: Results for held-out test set of LLL, with or without co-references

Considering results on raw data, our LM-kNN approach over-performs the other ones and produces high results for the extraction task. Besides this technique, it is interesting to note that the other ML approaches (libSVM and Random Forest in Table 1) exploiting our bag-of-lemmas representation and our 3-class modeling, despite their simplicity, also perform well compared with state-of-the-art techniques.

## 4.3 Discussion

The use of (deep) syntactic information for relation extraction seems very attractive at first sight. Indeed, many recent studies rely on the syntactic path between the two entities on which either hand-written extraction rules are applied (Fundel *et al.*, 2007, for example) or specially suited machine learning algorithms like string kernels or walk-weighted subsequence kernels for SVM are trained (Kim *et al.*, 2010). The results obtained are promising, yet, as we pointed it out before, they are highly dependent on the availability and the quality of the syntactic analysis (see (Fayruzov *et al.*, 2008)). For instance, the f-measure of (Kim *et al.*, 2010)

| System | Precision | Recall | F-measure |
|---|---|---|---|
| systems using raw data | | | |
| (Hakenberg *et al.*, 2005) | 50.0 | 53.8 | 51.8 |
| (Kim *et al.*, 2010) | 68.5 | 68.5 | 68.5 |
| (Fundel *et al.*, 2007) | 68 | 78 | 72 |
| LM-kNN | 67.1 | 87 | 75.8 |
| systems using manually annotated data | | | |
| (Popelínský & Blaťák, 2005) | 37.9 | 55.5 | 45.1 |
| (Riedel & Klein, 2005) | 60.9 | 46.2 | 52.6 |
| (Kim *et al.*, 2010) | 79.3 | 85.1 | 82.1 |

Table 2: Results for held-out test set of LLL, without co-references

declines by 15% when moving from a manual, perfect syntactic annotation to an automatic one (see Table 2). Thus, even for English and clean text, for which the parsing is the most reliable, using shallow linguistic approaches like our LM-kNN approach still yields better results for a fully automatic approach. In application domains with non formal text like speech transcripts or under-resourced languages, shallow linguistic information is clearly a more reliable resource than parsing.

# 5 Conclusion

In this paper, we have presented and experimented several systems, that can be easily implemented, to extract directed Protein-Protein Interactions in bio-medical texts. We have shown that modeling the PPI extraction task as a classification problem and simply using shallow linguistic information is sufficient to reach good results. Moreover, we have proposed a simple yet very efficient relation extraction system, LM-kNN, based on language modeling, which better takes the specifics of the task and data into account. The results, evaluated on a publicly available dataset, underlined the interest of using shallow linguistic information and our new LM-kNN method yielded the best known results.

This good result is very promising, and many perspectives are foreseen. From a technical point of view, it is possible to integrate these machine learning frameworks into an iterative process: newly retrieved relations are used as additional examples to re-train a system. Such approaches, like the one of (Hearst, 1992), as well as active learning techniques are of course straightforward for our lazy-learning approach. From an applicative point of view, our LM-kNN has to be tested over other relation extraction tasks. In particular, we foresee its use for the detection of relations in speech transcripts of sporting events. As it was previously said, shallow linguistic approaches is a necessity in such a context since the oral characteristics and the speech-to-text process make the use of deep linguistic analysis much more problematic.

# References

AIROLA A., PYYSALO S., BJÖRNE J., PAHIKKALA T., GINTER F. & SALAKOSKI T. (2008). A graph kernel for protein-protein interaction extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, p. 1–9, Columbus, Ohio, USA.

BREIMAN L. (2001). Random forests. *Machine Learning*, **45**(1), 5–32. doi:10.1023/A:1010933404324.

BUNESCU R. & MOONEY R. (2006). Subsequence kernels for relation extraction. *Advances in Neural Information Processing Systems*, **18**, 171–178.

CHANG C.-C. & LIN C.-J. (2001). *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

FAYRUZOV T., COCK M. D., CORNELIS C. & HOSTE V. (2008). The role of syntactic features in protein interaction extraction. In *Proceedings of the 17th Conference on Information and Knowledge Management (CIKM'08)*, p. 61–68, Napa Valley, CA, USA. doi:10.1145/1458449.1458463.

FAYRUZOV T., COCK M. D., CORNELIS C. & HOSTE V. (2009). Linguistic feature analysis for protein interaction extraction. *BMC Bioinformatics*, **10**. doi:10.1186/1471-2105-10-374.

FUNDEL K., KUFFNER R. & ZIMMER R. (2007). RelEx - relation extraction using dependency parse trees. *Bioinformatics*, **23**(3), 365–371. doi: 10.1093/bioinformatics/btl616.

GIULIANO C., LAVELLI A. & ROMANO L. (2006). Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, p. 401–408, Trento, Italy.

GOADRICH M., OLIPHANT L. & SHAVLIK J. (2005). Learning to extract genic interactions using gleaner. In *Proceedings of the 4th Learning Language in Logic Workshop (LLL05)*, p. 62–68, Bonn, Germany.

HAKENBERG J., PLAKE C., LESER U., KIRSCH H. & REBHOLZ-SCHUHMANN D. (2005). Lll'05 challenge: Genic interaction extraction - Identification of language patterns based on alignement and finite state automata. In *Proceedings of the 4th Learning Language in Logic Workshop (LLL05)*, p. 38–45, Bonn, Germany.

HALL M., FRANK E., HOLMES G., PFAHRINGER B., REUTEMANN P. & WITTEN I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations*, **11**(1), 10–18.

HEARST M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING'92)*, p. 539–545, Nantes, France.

KATRENKO S., MARSHALL M. S., ROOS M. & ADRIAANS P. (2005). Learning biological interactions from medline abstracts. In *Proceedings of the 4th Learning Language in Logic Workshop (LLL05)*, p. 53–58, Bonn, Germany.

KIM S., YOON J., YANG J. & PARK S. (2010). Walk-weighted subsequence kernels for protein-protein interaction extraction. *BMC Bioinformatics*, **11**. doi:10.1186/1471-2105-11-107.

LI M.-H., LIN L., WANG X.-L. & LIU T. (2007). Protein-protein interaction site prediction based on conditional random fields. *Bioinformatics*, **23**(5), 597–604. doi:10.1093/bioinformatics/btl660.

NÉDELLEC C. (2005). Learning language in logic – Genic interaction extraction challenge. In *Proceedings of the 4th Learning Language in Logic Workshop (LLL05)*, p. 31–37, Bonn, Germany.

NEY H., ESSEN U. & KNESER R. (1994). On structuring probabilistic dependencies in stochastic language modelling. *Computer Speech and Language*, **8**, 1–38.

PHUONG T. M., LEE D. & LEE K. H. (2003). *Learning rules to extract protein interactions from biomedical text*, In *Advanced in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science*, volume 2637, p. 148–158. Springer Verlag. doi:10.1007/3-540-36175-8_15.

POPELÍNSKÝ L. & BLAŤÁK J. (2005). Learning genic interactions without expert domain knowledge: Comparison of different ilp algorithms. In *Proceedings of the 4th Learning Language in Logic Workshop (LLL05)*, p. 59–61, Bonn, Germany.

PYYSALO S., AIROLA A., HEIMONEN J., BJÖRNE J., GINTER F. & SALAKOSKI T. (2008). Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, **9**(Suppl 3). doi:10.1186/1471-2105-9-S3-S6.

RIEDEL S. & KLEIN E. (2005). Genic interaction extraction with semantic and syntactic chains. In *Proceedings of the 4th Learning Language in Logic Workshop (LLL05)*, p. 69–74, Bonn, Germany.

SUN C., LIN L., WANG X. & GUAN Y. (2007). *Using maximum entropy model to extract protein-protein interaction information from biomedical literature*, In *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues, Lecture Notes in Computer Science*, volume 4681, p. 730–737. Springer Verlag. doi:10.1007/978-3-540-74171-8_72.

XIAO J., SU J., ZHOU G. & TAN C. (2005). Protein-protein interaction extraction: A supervised learning approach. In *Proceedings of the 1st International Symposium on Semantic Mining in Biomedicine (SMBM 2005)*, p. 51–59, Hinxton, Cambridgeshire, UK.