Learning context-free grammars to extract relations from text

Georgios Petasis[1] and Vangelis Karkaletsis[1] and Georgios Paliouras[1] and Constantine D. Spyropoulos[1]

Abstract. In this paper we propose a novel relation extraction method, based on grammatical inference. Following a semi-supervised learning approach, the text that connects named entities in an annotated corpus is used to infer a context free grammar. The grammar learning algorithm is able to infer grammars from positive examples only, controlling overgeneralisation through minimum description length. Evaluation results show that the proposed approach performs comparable to the state of the art, while exhibiting a bias towards precision, which is a sign of conservative generalisation.

1 INTRODUCTION

Relation extraction is the task of identifying the relations that hold between interesting entities in text data. Being a challenging subtask of information extraction, it extracts the knowledge required to move from named entity recognition to data interpretation and understanding. Thus, it has been one of the main areas of research in the field of computational linguistics. Initial attempts were mainly rule based [1] involving manually constructed rules, based on the results of syntactic analysis. Current research focuses mostly on the use of machine learning techniques. Supervised techniques have been shown to be effective for the task ([2];[3];[4]), while several approaches employ semi-supervised or unsupervised learning ([5];[6];[7];[8];[9];[10]), using also the Web as a corpus.

In this paper, a supervised machine learning approach is proposed. Assuming the existence of a named entity recogniser (NERC), the proposed approach extracts binary relations between named entities already identified in texts. Operating at the sentence level, a context-free grammar (CFG), which captures the patterns connecting related entities, is inferred from positive examples only. The eg-GRIDS ([11];[12]) grammatical inference algorithm that is used to learn the grammar, infers a CFG from positive examples only. The need for negative feedback to control overgeneralisation, is eliminated through the use of minimum description length (MDL) [13].

The main aim of this paper is to examine the suitability of grammatical inference for the task of relation extraction. A large part of the work done in the field exploits the results of syntactic analysis, along with statistical information obtained from large corpora, to acquire/generalise rules/patterns in order to perform relation extraction ([14];[15];[9];[17]). Starting from a parse tree that can been generalised by merging tree nodes [9], or from word sequences that can be converted into rules by exploiting information from parse trees [16], various heuristics have been proposed to drive the generalisation process and control the level of generalisation performed (in order to avoid over/under-generalisation). A general-purpose grammatical inference algorithm, on the other hand, already includes the required strategy for guiding generalisation along with the required stopping criteria. In addition, a grammatical inference algorithm is expected to be able to capture the syntactic structure of the language, minimising the need to perform syntactic analysis, making the approach more suitable for thematic domains where syntactic analysis exhibits reduced performance or for languages where the required processing resources are not available.

The criteria that have led to the selection of eg-GRIDS for relation extraction include its ability to infer grammars from positive examples only, the diversity of the search strategies implemented and the performance of the algorithm in the Omphalos context-free language learning competition [18]. Evaluation results show that the proposed method performs comparatively to the state of the art, while exhibiting a bias towards high precision, which can be attributed to the conservative generalisation approach of eg-GRIDS. Novel aspects of the proposed method include the ability to learn grammars autonomously, without relying on the availability of processing resources like part-of-speech taggers or syntax analysers. For example, many existing approaches use the results of syntactic analysis to generalise an initial hypothesis, or use syntax trees as the initial hypothesis to be generalised through node merging. Our approach eliminates these dependencies on processing resources, at the cost of extracting the required knowledge from the data directly. Thus, instead of applying heuristics to adapt a general-purpose grammar, such as the grammar of a conventional syntax analyser, into a specialised grammar for relation extraction, our approach concentrates on extracting the target grammar directly. Equally important is also the fact that the proposed approach does not rely on any sort of negative feedback, either direct, like the requirement for negative examples or irrelevant documents, or indirect, i.e. by assuming all data not positively annotated as negative examples, to control the level of generalisation performed. The advantage of not requiring additional resources and negative information increases the portability of the proposed approach not only to new thematic domains and languages, but perhaps also to other learning paradigms, like for example minimally supervised approaches: requiring only a limited amount of seed positive examples (or rules), the aim is to learn a target grammar through bootstrapping with respect to a corpus.

The rest of the paper is organised as follows: section 2 presents the proposed approach and introduces the grammatical inference algorithm, followed by an evaluation presented in section 3. Section 4 discusses work related to the presented approach, while section 5 concludes and outlines plans for future research.

[1] Software and Knowledge Engineering Laboratory, National Centre for Scientific Research – N.C.S.R. “Demokritos”, Athens, Greece, e-mails: {petasis, vangelis, paliourg, costass}@iit.demokritos.gr

2 EXTRACTING RELATIONS

In this section the proposed approach for relation extraction, using the eg-GRIDS algorithm is presented. More details about eg-GRIDS can be found at [11], [12].

2.1 Extracting relations

The task of extraction for a single relation type can be described as follows: Given a data set D and an n-ary relation *Rel* with arguments $X, Y, \ldots, Z$, find all instances $x \epsilon X, y \epsilon Y, \ldots, z \epsilon\ Z\ (x, y, z \epsilon D)$, such as $Rel(x, y, \ldots, z)$ holds [19]. The approach presented in this paper concentrates on extracting binary relations from textual corpora, by trying to capture the linguistic evidence in the text that connects two related entities.

In the training phase the method requires a set of training examples as input. The required examples can be easily obtained, if a corpus annotated with entities and relations between these entities is assumed. Each training example comprises the set of tokens (words) that lie between two related named entities x, y (including punctuation marks), and is labelled by the relation type $Rel(X, Y)$. If any named entity w is contained in such a training word sequence, all the tokens that constitute the named entity are replaced by the type of the entity (i.e. if "United States" is found, it is replaced with country), as the main focus is on capturing the information between entities and not the linguistic structure of entities, which is the task of a named entity recogniser.

From the set of training examples a set of context-free grammars is inferred, one for each relation type found in the training examples. The result of the training phase is a set of context-free grammars, one for each relation that can be extracted. Each context-free grammar is then converted into a classifier with the help of Boost.Xpressive C++ library [20]. Such a classifier returns true if the content between two entities can be parsed by the grammar and false otherwise.

2.2 The eg-GRIDS algorithm

The eg-GRIDS grammatical inference algorithm learns context-free grammars solely from positive example sets. Utilising a limited set of generalisation operations, eg-GRIDS follows an iterative approach in order to generalise an initial "flat" grammar extracted from the (positive) training examples. In each iteration, candidate grammars are scored according to the MDL heuristic, while search in the space of possible grammars can be directed by various search strategies (such as beam search or genetic evolution) and heuristics, which try to reduce training time through the detection of specific grammatical structures.

2.2.1 A bias towards "simple" grammars

As eg-GRIDS uses no negative evidence, an additional criterion is needed to direct the search through the space of context-free grammars and avoid overly general grammars. The approach of *minimum description length (MDL)* has been adopted in eg-GRIDS, which directs the search process towards grammars that are compact, i.e., ones that require few bits to be encoded, while at the same time they encode the example set in a compact way, i.e. few bits are required to encode the examples using the grammar. Assuming a context-free grammar G and a set of examples (sentences) T that can be recognised (parsed) by the grammar G, the total description length of a grammar, henceforth *model description length* abbreviated as *ML,* is the sum of two independent lengths:

- The grammar description length (*GDL*), i.e. the bits required to encode the grammar rules and transmit them to a recipient who has minimal knowledge of the grammar representation, and
- The derivations description length (*DDL*), i.e. the bits required to encode and transmit all examples in the set T as encoded by grammar G, provided that the recipient already knows G.

The first component of the *ML* directs the search away from the sort of trivial grammar that has a separate rule for each training sentence, as this grammar will have a large *GDL*. However, the same component leads to another sort of trivial grammar, a grammar that accepts all sentences (i.e. the most general grammar, "$S \rightarrow S\ T;\ T \rightarrow (any\ terminal \mid e)$"). In order to avoid this, the second component estimates the *derivation power* of the grammar, by measuring the way the *training examples* are generated by the grammar, and helps to avoid overgeneralisation by penalising general grammars. The higher the derivation power of the language, the higher its *DDL* is expected to be. The initial overly specific grammar is trivially best in terms of *DDL*, as usually there is a one-to-one correspondence between the examples and the grammar rules, i.e. its derivation power is low. On the other hand, the most general grammar has the worst score, as it involves several rules in the derivation of a single sentence, requiring substantial effort to track all the rules involved in the generation of the sentence.

2.2.2 Architecture of eg-GRIDS and the learning operators

The architecture of eg-GRIDS is summarised in Figure 1. eg-GRIDS uses the training sentences in order to construct an *initial, "flat" grammar*. This initial grammar is constructed by simply converting each one of the training examples into a grammar rule[2]. As a result, the number of initial rules corresponds to the number of training examples. This initial grammar is overly specific, as it can recognise only the sentences contained in the training set. After the initial grammar has been created, eg-GRIDS generalises this initial grammar, using one of the two available iterative search processes: beam or genetic search. Both search strategies utilise the same search operators in order to produce more general grammars. Currently, eg-GRIDS supports five search operators:

Merge NT: merges two non-terminal symbols into a single symbol, thereby replacing all their occurrences in all rules with the new symbol.

Create NT: creates a new non-terminal symbol X, which is defined as a sequence of two or more existing non-terminal symbols. X is defined as a new production rule that decomposes X into its constituent symbols.

Create Optional NT: duplicates a rule created by the "Create NT" operator and appends an existing non-terminal symbol at the end of the body of the rule, thus making this symbol optional.

Detect Center Embedding: aims to capture the center embedding phenomenon. This operator tries to locate the most frequent four-gram[3] of the form "A A B B". Once such a four-gram is located, the operator creates a new non-terminal symbol X as the operator "Create NT" would have done. However, *assuming* that this four-gram was created through center embedding involving symbol X, this operator additionally creates a new production rule of the form "$X \rightarrow A\ A\ X\ B\ B$" and replaces all symbol sequences that match the pattern "$A{+}\ X?\ B{+}$" with X.

Rule Body Substitution: examines whether the body of a production rule R is contained in bodies of other production rules.

[2] The body of each rule is a sequence of non-terminal symbols, as each terminal is mapped initially to a unique non-terminal.

[3] Since bigrams and trigrams are quite common (frequent) structures and their presence can be attributed to a large number of phenomena, four-grams are assumed to be the smallest n-grams that indicate possible existence of center embedding.

In such a case, every occurrence of the body of rule *R* in other rule bodies is replaced by the head of rule *R*.

The five operators create grammars that have either the same or greater expressiveness than their parent grammar. As the operators never remove rules from a grammar, the resulting grammars have at least the same coverage as the parent grammar, i.e. they can recognise at least the same set of sentences.

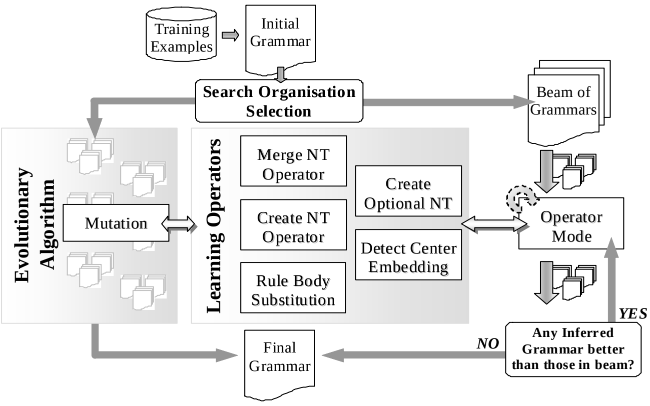

Figure 1: The architecture of the eg-GRIDS algorithm.

3 EVALUATION

For the purposes of the evaluation, annotated corpus from the BOEMIE research project was used. The corpus contained 800 HTML pages, retrieved from various sites of athletics associations like IAAF[4], EAA[5] and USATF[6], containing pages with news, results and athlete's biographies.

All pages have been manually annotated, according to a semantic model capturing information about athletes and their participations in sports competitions, held under official competitions. This semantic model formed also the basis for annotating the corpus with relations. A named entity recogniser developed in the context of the BOEMIE project was applied to the corpus, to identify named entities related to the athletics domain. The recogniser uses Conditional Random Fields [21], and exhibits precision of 90 %, and recall approaching 86 %. Once the corpus has been annotated with named entities, entities representing the same real objects or events were identified through matching, and associated with the entities of the semantic model. Having an alignment between identified entities and the semantic model, relations in the semantic model can be projected on the corpus, providing an initial annotation of binary relations between the identified entities. As a next step in the preparation of the data, the relations involving person names and person properties like gender, age, nationality, performance and ranking were manually verified and corrected where necessary.

The evaluation was limited to relations occurring within sentence boundaries, in order to keep the complexity of the grammars to be learned, and thus the required time to learn them, at tractable levels. This is the main reason also for considering only relations involving names and properties related to athletes, as their vast majority does not cross sentence boundaries, in contrast to relations involving athletes and sport competitions or athletic events they have participated in. Thus, as a final step, relations crossing sentence boundaries were removed from the corpus, producing a corpus with 8.497 relations involving person names and person properties.

From this corpus, a set of 8.497 training examples was created. To reduce data sparseness, word stems were used instead of the actual words. Each training example contained all word stems and punctuation symbols found in the corpus between two related entities, in the order that they appear in the corpus. Each entity found into the training example was replaced by its entity type, while each example was labelled with the entity types of the two related entities. An example of a sentence annotated with named entities is shown in Figure 2, while the generated training examples are shown in Figure 3.

Kenya=[country]'s **Richard Limo**=[name] the World **5000m**=[sport_name] champion (eventual **third**=[ranking] **26:50.20**=[performance]) came the nearest during the first 300m of the lap, until in the finishing straight, **Ethiopia**=[country]'s Olympic **bronze**=[ranking] **Assefa Mezegebu**=[name] started a drive to the line which took **second**=[ranking] place (**26:49.90**=[performance]).

Figure 2: A sample sentence annotated with named entities.

Evaluation was performed through 10 fold-cross validation, and performance was measured in terms of precision, recall and F-measure. In each fold, one grammar per relation type was inferred from 9/10 of the training examples. The unseen 1/10 of the examples held for evaluation was parsed by all inferred grammars: if an example was parsed correctly only by the grammar corresponding to the correct relation type, the example was considered correct. In all other cases, including the case where an example was parsed by more than one of the learned grammars, the example was considered a failure. The obtained performance results are shown in Table 1.

Word stems	Relation label
' s	name-country
the world entity:sport_name champion (eventual	name-ranking
the world entity:sport_name champion (eventual entity:ranking	name-performance
' s	name-country
start a drive to the line which take	name-ranking
start a drive to the line which take entity:ranking place (	name-performance

Figure 3: Training examples extracted from the sample sentence of Figure 2.

	Precision	Recall	F-measure
Name-Ranking	95.05 %	54.07 %	68.57 %
Name-Performance	92.14 %	49.26 %	64.17 %
Name-Country	98.85 %	88.88 %	93.58 %
Name-Gender	99.21 %	79.17 %	88.00 %
Name-Age	100.00 %	98.11 %	99.04 %
Overall	**96.48 %**	**65.96 %**	**78.32 %**

Table 1: Performance results.

Evaluation results suggest that the proposed approach performs well in comparison to the state of the art, despite the difficulties of comparing results obtained on different corpora. For example, in [9], the presented approach, expanding on a basis of 55 manually constructed seed rules, exhibits precision around 88 % with 43 % recall on 1032 news reports on Nobel prizes from New York Times, BBC and CNN.

The fact that our approach uses as input only word stems has two interesting implications: (a) if an example contains a stem that has not been seen before, this example will always be classified as a failure, as it cannot be parsed by any grammar and (b) any generalisation can only be attributed to the successful operation of eg-GRIDS in forming the correct syntactic abstractions, in order to allow the use of "similar" stems instead of a specific stem. One

[4] International Association of Athletics Federations – http://www.iaaf.org/.
[5] European Athletics Association – http://www.european-athletics.org/.
[6] USA Track and Field – http://www.usatf.org/.

easy answer to (a), followed by numerous approaches (e.g. [16]) in the literature, is to add another level of abstraction over words, such as part-of-speech tags. The fact that the presented approach does not make use of such an abstraction layer, allows us to obtain an estimate of the generalisation achieved solely by the grammatical inference algorithm in use. For this reason, the same experiment was repeated with a slight change: duplicate entries were removed from the training example set, making all training examples unique. This reduced the training example set by almost 2/3, but ensured that all examples used for evaluation had never been seen during training. Again 10 fold-cross validation was used, and the evaluation results are shown in Table 2.

	Precision	Recall	F-measure
Name-Ranking	50.04 %	6.79 %	11.90 %
Name-Performance	67.16 %	11.87 %	20.13 %
Name-Country	100.00 %	16.05 %	27.20 %
Name-Gender	74.83 %	7.04 %	12.73 %
Name-Age	80.00 %	47.12 %	55.00 %
Overall	**67.58 %**	**10.46 %**	**18.09 %**

Table 2: Performance results on unique training example set.

Despite the fact that the results of Table 2 are a pessimistic approximation (since examples containing unknown words with respect to the training examples have not been eliminated), eg-GRIDS managed to achieve a generalisation of about 10 pp in terms of recall, which is impressive considering that this involves word usages in an ordering never observed during training, even if the loss in precision approaches 29 pp.

Regarding execution time during grammar learning, the eg-GRIDS algorithm is able to converge to a final grammar within a few minutes (from 5 to 15 minutes in most cases) when learning from the complete training example set in the evaluation experiment performed first. However, converting the learned context-free grammar into a classifier (through the use of the template-based Boost.Xpressive C++ library) required considerable amounts of compilation time[7], in the range of 45 to more than 60 minutes per grammar.

4 RELATED WORK

To our knowledge, there is very little work on relation extraction with grammatical inference. In [14] a semi-automated approach is presented, which exploits the results of corpus analytics (mainly concordances of verbs) to propose patterns. These patterns, after being validated by an expert, are converted into a set of finite state automata. Similarly, in [15] automata are again used, created from manually constructed patterns. Both approaches however operate on syntactic trees obtained through parsing and involve manually or semi-automatically constructed patterns for extracting relations.

On the other hand, there are some approaches that exhibit some resemblance in the sense that they try to generalise extracted patterns/rules [16], or modify extraction rules by applying operators similar to the ones employed by eg-GRIDS [9]. LearningPinocchio [17] has been built upon the LP^2 algorithm [16], which creates an initial set of rules from positive examples that are generalised by exploiting the results of linguistic analysis/shallow syntactic parsing to remove constraints from the rules. Overgeneralisation is controlled through negative examples, obtained automatically from the corpus, under the assumption that everything not marked as a positive example is a negative one. Following a similar approach, DARE [9] starts with a minimal number of seed rules which are used to annotate a corpus. Having as input syntax trees, DARE follows a bottom-up approach to obtain more general rules by merging nodes of the syntax trees of sentences, an operation that is also part of eg-GRIDS, as one of its generalisation operators. Overgenaralisation is controlled by trying to maximise rule matches in relevant documents while maintaining a small number of matches in irrelevant documents. Our approach differs from these two by not depending on syntactic analysis (used either as a starting point for extracting rules in DARE or for guiding generalisation in LearningPinocchio). Our method also uses MDL for controlling overgeneralisation, thus eliminating the need for negative feedback.

5 CONCLUSIONS

Relation extraction methods typically involve the acquisition of extraction rules and grammars: from Hearst patterns [5] that try to detect hierarchical relation such as hypernyms, to complex lexico-syntactic grammars [9] aiming at extracting *n*-ary relations with *n > 2*. Being mainly supervised or semi-supervised methods, they frequently combine syntax trees obtained through syntactic analysis with heuristics based on various statistical measures, in order to generalise an initial hypothesis formed from the training data. In an attempt to ease the requirements posed by such approaches we have examined the suitability of a general purpose grammatical inference algorithm to the task, aiming to evaluate its suitability in replacing both the need for syntactic analysis as well as the heuristics required to guide the generalisation process. The proposed approach has been evaluated with the help of a manually annotated corpus and the obtained evaluation results suggest that the approach performs comparatively to the state of the art, without requiring additional resources such as syntactic analysis or part-of-speech tagging. In addition, the fact that the proposed approach does not involve any abstraction other than the generalisation performed by the grammatical inference algorithm, allowed us to get an estimate of the degree of generalisation that can be achieved by the algorithm. This was measured to be at least 10 pp accompanied with a degradation in precision of about 29 pp.

Since the obtained results are satisfactory, it seems interesting to try to eliminate also the requirement of utilising a manually annotated corpus. Unsupervised approaches have attracted significant research interest, as manual annotation is a time-consuming and resource needing process. Following recent advances in the field, the proposed approach can be adapted to accept a set of seed context-free grammars, with each one containing only a few rules targeting a specific relation type. Utilising a bootstrapping procedure the system may try to generalise these seed grammars with respect to a corpus of documents relevant to the domain of interest.

Acknowledgements

This work has been partially funded by the BOEMIE Project, FP6-027538, 6th EU Framework Programme.

6 REFERENCES

[1] 7th Message Understanding Conference (MUC-7), April, 1998. http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_toc.html

[2] CoNLL, 2008: http://www.yr-bcn.es/conll2008/

[3] Carreras X. and Màrquez L., “Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling”, *In Proc. of the Ninth Conference on*

[7] The experiment was conducted on a PC running Windows Vista (64bit), with an Intel 6700 processor and 4 GB or RAM. The compiler used was MS VC++ 2005.

Natural Language Learning (CoNLL-2005), June 29 – 30, Michigan, USA, 2005. http://www.lsi.upc.edu/~esrlconll/st05/st05.html

[4] Carreras X. and Màrquez L., “Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling”, *In Proc. of the Eighth Conference on Natural Language Learning (CoNLL-2004)*, Workshop of HLT/NAACL 2004, May 6 – 7, Boston, MA, USA, 2004. http://www.lsi.upc.edu/~srlconll/st04/st04.html

[5] Hearst M., “Automatic acquisition of hyponyms from large text corpora”, *In Proc. of the 14th International Conference on Computational Linguistics (COLING-1992)*, 1992.

[6] Davidov D., Rappoport A., and Koppel M., “Fully Unsupervised Discovery of Concept-Specific Relationships by Web Mining”, *In Proc. of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 232 – 239, Prague, Czech Republic, June, 2007. http://www.aclweb.org/anthology/P/P07/P07-1030

[7] Brody S., “Clustering Clauses for High-Level Relation Detection: An Information-theoretic Approach”, *In Proc. of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 448 – 455, Prague, Czech Republic, June, 2007. http://www.aclweb.org/anthology/P/P07/P07-1057

[8] Bunescu R., and Mooney R., “Learning to Extract Relations from the Web using Minimal Supervision”, *In Proc. of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 576 – 583, Prague, Czech Republic, June, 2007. http://www.aclweb.org/anthology/P/P07/P07-1073

[9] Xu F., Uszkoreit H., Li H., “A Seed-driven Bottom-up Machine Learning Framework for Extracting Relations of Various Complexity”, *In Proc. of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 584 – 591, Prague, Czech Republic, June, 2007. http://www.aclweb.org/anthology/P/P07/P07-1074

[10] Rosenfeld B., and Feldman R., “Using Corpus Statistics on Entities to Improve Semi-supervised Relation Extraction from the Web”, *In Proc. of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 600 – 607, Prague, Czech Republic, June, 2007. http://www.aclweb.org/anthology/P/P07/P07-1076

[11] Petasis G., Paliouras G., Spyropoulos C. D. and Halatsis C., “eg-GRIDS: Context-Free Grammatical Inference from Positive Examples using Genetic Search”. *In Proc. of the 7th International Colloquium on Grammatical Inference (ICGI 2004)*, Lecture Notes in Artificial Intelligence 3264, pp. 223 – 234, Springer, 2004.

[12] Petasis G., Paliouras G., Karkaletsis V., Halatsis C., and Spyropoulos C. D., “e-GRIDS: Computationally Efficient Grammatical Inference from Positive Examples”. *GRAMMARS*, (7), pp. 69 – 110, 2004. (http://grammars.grlmc.com/special.asp)

[13] Rissanen J., “Stochastic Complexity in Statistical Inquiry”, *World Scientific Publishing Co*, Singapore, 1989.

[14] Pustejovsky J., Castano J., Zhang J., Cochran B., and Kotecki M., “Robust relational parsing over biomedical literature: Extracting inhibit relations” *In Pacific Symposium on Biocomputing*, 2002. http://citeseer.ist.psu.edu/527763.html

[15] Leroy G. and Chen H., “Genescene: An Ontology-enhanced Integration of Linguistic and Co-occurrence based Relations in Biomedical Texts” *In Journal of the American Society for Information Systems and Technology (JASIST)*, 56 (5), 457 – 468, March 2005. http://beta.cgu.edu/Faculty/leroyg/Leroy-JASIST-2005.pdf

[16] Ciravegna F., “Adaptive information extraction from text by rule induction and generalization”, *In Proc. of the 17th International Joint Conference on Artificial Intelligence (IJCAI 2001),* 2001.

[17] Ciravegna F., and Lavelli A., “LearningPinocchio: Adaptive Information Extraction for Real World Applications”, *In Journal of Natural Language Engineering*, 10 (2), 2004.

[18] Starkie B., Coste F., van Zaanen M. “The Omphalos Context-free Grammar Learning Competition”, *In Grammatical Inference: Algorithms and Applications; Proc. of 7th International Colloquium on Grammatical Inference (ICGI 2004)*, vol. 3264 of LNCS/LNAI, pp. 16 – 27, Springer-Verlag, 2004.

[19] Katrenko S., and Adriaans P., “Learning Relations from Biomedical Corpora Using Dependency Tree Levels”, *In Proceedings of the Fifteenth Dutch-Belgian Conference on Machine Learning (Benelearn)*, Ghent, Belgium, May 12, 2006. http://staff.science.uva.nl/~katrenko/katrenko_adriaans.pdf

[20] http://boost-sandbox.sourceforge.net/libs/xpressive/doc/html/index.html

[21] Lafferty J., McCallum A., and Pereira F., “Conditional random fields: Probabilistic models for segmenting and labeling sequence data”, *In Proc. of ICML*, pp.282 – 289, 2001.