# Description of the LISN system for extracting terms

Thierry Hamon[1,2,*]

[1]Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400, Orsay, France
[2]Université Sorbonne Paris Nord, 99 avenue J-B Clément, 93430 Villetaneuse, France

**Abstract**

This paper describes the LISN system which addresses the terminology extraction task within the challenge DETECH 2026. The system architecture is composed of two sub-systems. The first one focuses on a rule-based term extraction approach relying on the state-of-the-art term extractor YATEA. In the second sub-system, term extraction is considered as a sequential tagging. Two different representations of terms are then used to train a neural network model. The overall system is tested on the two corpora of the challenge (mental health and Parkinson disease). The results show that the symbolic approach achieved better results than the machine learning approach. However, we observe that when tagging the term contexts, rather than the term components, leads to promising results.

**Keywords**

automatic term extraction, terminology, rule-based ATE, sequential tagging

## 1. Introduction

Terminology provides important information in text-based domain-specific applications such text mining or translation, thus providing an access to the knowledge of the domain. However, terminology, even if it exists for a given domain, may not be fully tuned for the texts to be processed or for the application. Hence, automatic term extraction becomes helpful for extending the terminology coverage or for building the terminology associated to the working text corpus.

Even if automatic term extraction (ATE) has be tackled for decades, it remains a challenging task since the nature of the terms may be difficult to define. Hence, terms can be described as lexical items that refer to concepts of a domain [1]. Extracting terms requires to identify their unithood, i.e. the strength of the term components, and their termhood, i.e. the strength of the relation between the terms and the domain [2]. This question is addressed by the DETECH 2026 challenge [3].

In the paper, we describes the system developed for the ATE task. After presenting related work (section 2), we describe the task and the available data in Section 3. In Section 4, we detail the system architecture and the methodology. The results on the training and test sets are presented and discussed in Section 5. Then, we conclude with final remarks and present research perspectives (Section 6).

## 2. Related Work

Often, term extraction methods rely on hybrid approaches. First, linguistic characteristics of terms are taken into account in rule-based approaches to chunk the texts and extract term candidates [4]. More recently, thanks to the development and improvement of syntactic parsers, dependency parsers have been used for the term extraction [5].

If linguistic approaches for term extraction from texts have been well defined for decades, they still have difficulties to identify relevant terms among the huge amount of term candidates: indeed, terms and other irrelevant noun phrases extracted may have similar linguistic characteristics. Moreover, due

*International Workshop on DEfinition and Term Extraction Challenge (DETECH) 2026, June 24, 2026, Zadar, Croatia*
*Corresponding author.
thierry.hamon@lisn.fr (T. Hamon)
https://perso.lisn.upsaclay.fr/hamon/ (T. Hamon)
0000-0002-1521-4875 (T. Hamon)

to an increasing size of the corpora, users must check an important set of term candidates and often have some difficulty to identify relevant terms among all the noun phrases extracted [6].

After the extraction step, statistical approaches can be applied to compute the termhood of the extracted noun phrases regarding their occurrences within texts. To achieve this, statistical approaches tackle the issue by focusing on the filtering of lexical units: statistical metrics or contextual information on the noun phrases are used for (i) filtering out n-grams or ranking term candidates, or (ii) learning a model to classify lexical units as term or non-terms.

In the first case, many statistical metrics have been proposed for selecting or ranking term candidates. The frequency of noun phrases has been the first commonly used ranking metric [7, 8, 9]. But such statistical information is not sufficient to fully capture the termhood [10], as it tends to decrease the recall, because many term candidates occur only once in the corpus [8, 11, 12], or the precision [13]. The length of noun phrases, defined as the number of words they contain, has been also considered as a clue to distinguish terms among all the noun phrases extracted from texts. Thus, Drouin [11] proposes rather to consider the inverted length of the terms: the longer the term, the less important it is. Such information combined with the term frequency tends to slightly increase the precision: term candidates containing one word or short noun phrases become preferred. On the contrary, the C-Value measure [14] has the purpose to promote longer term candidates. Even if its impact has not been clearly evaluated yet, the increase of the precision seems to remain low. Regarding these results, the contribution of the term length appears to be mainly corpus-dependent. The context of the candidate terms is also assumed to be helpful to identify the termhood of the extracted noun phrases. Thus, the NC-Value attempts to combine contextual information with the C-Value [13]. Termhood can also be defined as the semantic relatedness to a domain, represented as a vector of generic words occurring in the corpus [15]. This kind of model outperforms the NC-Value on a biomedical corpus while the both remain equivalent for keyphrase extraction.

More recently, statistical information is used for learning models for term filtering [16, 17, 18]. Thus, [10] proposes to combine contrastive and consensus metrics to model the relevance and the lexical cohesion of candidate terms for a given domain. The evaluation of the metrics on a tourism corpus, after the pruning of the term list, and comparison with frequency, show an increase of the precision yet at the expense of the recall. Word embeddings are also used for term extraction [19, 20, 21, 22]. For instance, [23] considered ATE as a binary classification of n-grams and used pre-trained word embeddings, such as BERT, in combination with previously discussed statistical measures. More generally, in a recent survey [24], neural models, in particular transformer-based models and LLMs, improve the performances of the ATE, and may gain by taking into account feature engineering-based systems.

Comparison of linguistic-based methods and machine learning approaches is not obvious: evaluation can be limited and each approach shows some weaknesses [18]. More recently, in a systematic review, [25] show that even if neural models have promissing performances, the proportion of recognized terms is still low and requires to still include domain expert for building a high-quality terminology. Moreover, the lack of explainability and transparency of such ATE systems, in particular the neural models, is a important drawback in sensitive contexts [24]. Challenges, like TermEval or DETECH 2026, contribute to improve our knowledge on term extraction approaches.

## 3. Experimental Settings

For the ATE task, the challenge DETECH 2026 provides two text corpora in English issued from the medical domain. Each text set focuses on diseases related to the gut–brain interplay: mental health and Parkinson. Texts are issued from PubMed[1] citations. The annotation guidelines[2] describe the building process of these corpora. Each file within the corpora contains several fields: SO (source), TI (title), AB (abstract) (see Figure 1). As a preliminary cleaning, we remove the first line in each text file, i.e. the SO field, which reduced the risk to extract erroneous terms.

[1] https://pubmed.ncbi.nlm.nih.gov/

[2] https://github.com/gmdn/DETECH2026/blob/main/Annotation_Protocol/DETECH_Annotation_Protocol_2003.pdf

```
SO  - Psychiatry Res. 2024 Jun;336:115914. doi: 10.1016/j.psychres.2024.115914. Epub
      2024 Apr 16.
TI  - Effects of antipsychotics on the gastrointestinal microbiota: A systematic
      review.
AB  - Antipsychotics (APs) have been increasingly prescribed for psychiatric disorders
      from schizophrenia to disruptive behavioral conditions. These drugs have been
      associated with considerable side effects, such as weight gain, and increasing
      evidence has also indicated that its use impacts gut microbiota (GM), although
      this connection is still little understood. To assess APs effects on the GM of
      patients starting or ongoing treatment, a systematic review was carried out in
      PubMed and Scopus databases. Twelve articles were considered eligible for the
      review, which investigated the effects of risperidone (5 studies), quetiapine
     (3), amilsupride (1), olanzapine (1), and unspecified atypical drugs (2). Eleven
     reported changes in GM in response to APs, and associations between the abundance
      of bacterial groups and different metabolic parameters were described by (...)
```

**Figure 1:** Example of PubMed citation used by the challenge DETEC 2026.

**Table 1**
Description of the two corpora. Information on unannotated texts are in parenthesis.

| | training | | | | test | |
|---|---|---|---|---|---|---|
| Corpus | Texts | Words | Term occ | Term type | Text | Words |
| mental health | 356 | 87,493 | 14,384 | 4,733 | 356 | 89,029 |
| parkinson | 184 (40) | 44,402 (10,065) | 6,449 | 1,957 | 165 (46) | 41,231 (11,591) |

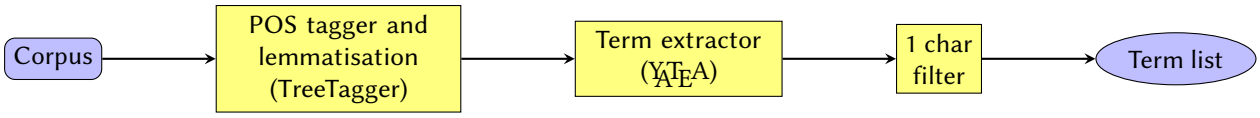


**Figure 2:** Architecture of the rule-based term extraction (Run 1).

Regarding the mental health topic, 356 texts were provided in the training set, and additional 356 texts provided in the test set. As for the Parkinson topic, the training set contained 184 texts while the test set has 165 texts. The Parkinson collection also included unannotated texts: 40 texts in training and 46 in test sets. Table 1 details the training and test data.

The objective of the ATE task is to extract, from the provided texts, terms which are relevant for the both domains: mental health and Parkinson. According to the annotation guidelines, terms can be multi-word or single-word terms, and must occur in the texts. Lexical units including logical connector as “and” and “or” are not considered as terms. Each occurrence of terms has to be identified in the texts. Table 1 provides the number of term occurrences and term types per training corpus.

## 4. Methodology

The architecture of our system is divided into two sub-systems. The first sub-system relies on a state-of-the-art term extractor (see Section 4.1), while in the second sub-system, term extraction is considered as a sequential tagging with two representations of terms for training a neural network model (see Section 4.2).

### 4.1. Rule-based Term Extraction

The rule-based system is based on state-of-the-art components (see Figure 2). Such rule-based system does not require learning step and is used similarly on the training and test corpora.

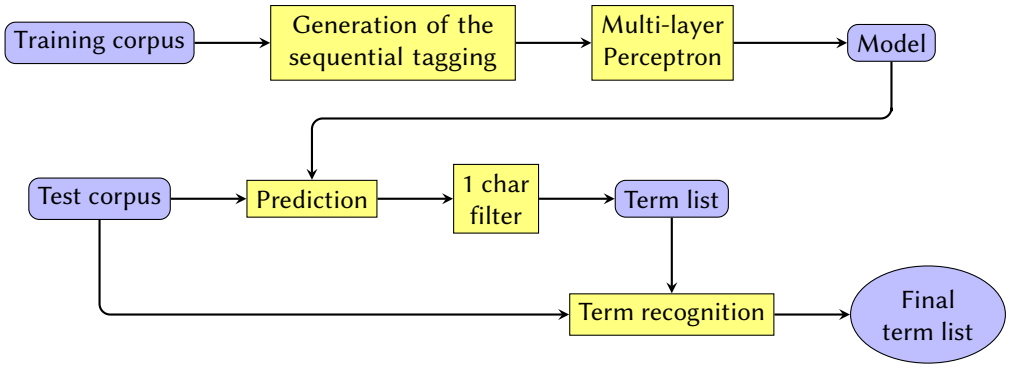


**Figure 3:** Architecture of the term extraction based on sequential tagging (Runs 2 and 3).

First, the part-of-speech tagger TreeTagger [26] is applied on the corpus in order to tokenize the texts in sentences and words and to provide the morphosyntactic information associated to words. Such information is required by the term extraction step. Second, we use the term extractor YaTeA[3] [27] on morphosyntactically analysed texts.

YaTeA performs shallow parsing of the POS-tagged and lemmatized texts by chunking them according to pre-defined syntactic frontiers (pronouns, conjugated verbs, typographic marks, etc.). This permits to identify noun phrases which are then syntactically analysed with parsing patterns in order to obtain parsed terminological entities. Each term is represented in a syntactic tree, and the sub-terms can also be considered as terms in the current configuration (e.g. *abundance of bacterial groups* leads to also extract *bacterial groups* and *abundance* from Figure 1). Single-word term can be also extracted.

The term extractor can propose various kinds of terms according to four configurations:

- multi-word terms which are not included in terms (e.g. *abundance of bacterial groups*)
- all multi-word terms (e.g. *abundance of bacterial groups* and *bacterial groups*)
- multi-word terms which are not included in larger terms and single-word terms (e.g. *abundance of bacterial groups* and *Antipsychotics*, but not *abundance* or *bacterial groups*)
- all multi-word and single-word terms (e.g. *abundance of bacterial groups*, *bacterial groups*, *abundance* and *Antipsychotics*)

The last configuration includes all the terms proposed by other configurations.

In order to be as close as possible to the DETECH 2026 annotation principles, we performed preliminary experiments on training corpora with the four configurations. We observed that extracting multi-word terms, which are not included in larger terms, and single-word terms, provides best performances, and termhood and unithood as defined by DETECH 2026 are better reflected.

Similarly, during these experiments, we also observed that terms with only one character lead to the extraction of too many erroneous terms. Therefore, even if the training term set contained such terms, we preferred to filter them out.

## 4.2. Sequential Tagging for Term Extraction

In this second sub-system, we considered the term extraction task as a sequential tagging task: each word of the texts is associated to a tag which triggers the term extraction.

Two tag sets at the word level have been investigated. First, we use the BILOU tag set. Each word of the texts is to be assigned to a class according to its position when considering terms: Beginning of the term (B-TERM), Inside a term (I-TERM), End of the term (L-TERM), Outside the terms (O), Unique word corresponding to one-word terms (U-TERM). Figure 4 presents an example of the BILOU annotation. As for the second tag set, the purpose was to focus on the context of the terms. In addition to outside-term

[3] http://search.cpan.org/~thhamon/Lingua-YaTeA/

```
Effects_O of_O antipsychotics_U-TERM on_O the_O gastrointestinal_B-TERM
microbiota_L-TERM :_O A_O systematic_O review_O ._O

Antipsychotics_U-TERM (_O APs_U-TERM )_O have_O been_O increasingly_O prescribed_O
for_O psychiatric_B-TERM disorders_L-TERM from_O schizophrenia_U-TERM to_O
disruptive_O behavioral_O conditions_O ._O
```

**Figure 4:** Tagging of the words according to the BILOU tag set.

```
Effects_I-2 of_I-1 antipsychotics_T on_I+1 the_I-1 gastrointestinal_T microbiota_T
:_I+1 A_I+2 systematic_O review_I-1 ._I-2

Antipsychotics_T (_M APs_T )_I+1 have_I+2 been_O increasingly_O prescribed_I-1 for_I-2
psychiatric_T disorders_T from_M schizophrenia_T to_I+1 disruptive_I+2 behavioral_O
conditions_O ._I-1
```

**Figure 5:** Tagging of the words according to the context tag set.

words (O) and term-component words (T), other word classes are related to the context of terms: Before the term at the $n$ position (I-n), After the term at the $n$ position (I+n), Between two terms (M). For this challenge task, we used the context size set to two words before and after the terms. Figure 5 presents an example of such context annotation.

We investigated two classifiers implemented in Python with Keras libraries: BiLSTM-CRF [28] and Multilayer Perceptron (MLP) [29] with 2 hidden layers including a ReLU activation and 0.2 dropout. The MLP classifier has the best performances on the training corpora. We created a model for each corpus and for each tag set. The features are related to the word structure (the inflectional form, prefixes and suffixes with 1 to 3 characters, presence of uppercased and lowcased characters, and presence of special characters and numbers) and to their context (inflectional forms of words within the 5-word windows on the left and on the right). The predictions were loosely interpreted: the term extraction is triggered by a non-O tag with the BILOU tagging, and by a T tag for context tagging. Similarly to the rule-based term extraction, one character terms are removed.

Our preliminary results also indicated that the type measures are better than the occurrence measures (micro-measures). It means that the system correctly extracts terms at the level of the whole corpus, but may miss some occurrences of these terms. To improve the recall, we look up for all the occurrences of all the extracted terms. Hence, the term set issued from the predictions is then used for term recognition.

## 5. Results and Discussions

We apply the two sub-systems on mental health and Parkinson corpora. The performances are evaluated at the occurrence level with micro-metrics ($\mu$-precision, $\mu$-recall and $\mu$-F1 measure) and at the document level (type metrics). The comparison with the reference is case-insensitive and has been carried out without lemmatisation.

Performances of the three runs on the test corpora are given in Table 2. When writing this paper, evaluation scripts were not available yet and performances on the training corpora could not be computed. The median results on all the participants runs is provided by the challenge. We note that the rule-based term extraction (Run 1) gives the best performance on the two corpora. $\mu$-F1 and type-F1 are 0.10 higher than the median results. As expected, the run 3 (BILOU tagging) gives the lowest results. These results show that rule-based system remains efficient on a ATE task. They also indicate that combination of context tagging with term recognition provides interesting performances, in particular for the recall at the occurrence level. At the type level, the BILOU tagging has the best precision. We assume that the run 3 needs more information for extending its coverage. To confirm this hypothesis,

**Table 2**
Results of the LISN system on the ATE task.

| | $\mu$Prec | $\mu$Rec | $\mu$F1 | type-Prec | type-Rec | type-F1 |
|---|---|---|---|---|---|---|
| | | | Mental health | | | |
| Run 1 (rule-based ATE) | **0.456** | **0.676** | **0.545** | 0.415 | **0.686** | **0.517** |
| Run 2 (context tagging) | 0.335 | 0.673 | 0.447 | 0.384 | 0.597 | 0.467 |
| Run 3 (BILOU tagging) | 0.397 | 0.447 | 0.420 | **0.516** | 0.368 | 0.43 |
| Median | 0.801 | 0.421 | 0.434 | 0.805 | 0.459 | 0.493 |
| | | | Parkinson | | | |
| Run 1 (rule-based ATE) | **0.454** | **0.664** | **0.539** | 0.426 | **0.684** | **0.525** |
| Run 2 (context tagging) | 0.357 | 0.634 | 0.456 | 0.398 | 0.56 | 0.465 |
| Run 3 (BILOU tagging) | 0.384 | 0.363 | 0.373 | **0.507** | 0.285 | 0.365 |
| Median | 0.621 | 0.395 | 0.415 | 0.622 | 0.408 | 0.428 |

we need to carry out further experiments with additional features.

Finally, contrary to the median results, we observe that the recall of rule-based term extractor (run 1) and context tagging (run 2) is better than the precision values. Moreover these recall values are higher than the median results (by more than 0.2). Since context tagging includes term recognition, it means that the two systems take advantage of the symbolic part of the extraction process. Thus, combining machine learning and symbolic approaches is an interesting perspective for improving ATE.

## 6. Conclusion and Future Work

While automatic term extraction remains an active field, the impact of the contribution of linguistic-based and machine learning based approaches increases with the organisation of challenges, thus permitting to compare various types of approaches. In this paper, we describe our contribution to the DETECH 2026 challenge. Our system is based on (i) an off-the-shelf rule-based term extraction (ii) a MLP classifier where term extraction is considered as sequential tagging. Compared to such low featured classifier, the rule-based system produces the best results. However, when a context tagging is used, the classifier also provides interesting results. In future work, we plan to investigate how a filtering step applied on the output of the rule-based term extractor may impact the results. As for the context tagging, we will use other features for the classifier, such as POS-tag, word embeddings from BERT, and other architectures.

## Declaration on Generative AI

The author has not employed any Generative AI tools.

## References


[1] K. Kageura, E. Marshman, Terminology extraction and management, in: M. O'Hagan (Ed.), The Routledge Handbook of Translation and Technology, 2019, pp. 61–77.

[2] K. Kageura, B. Umino, Methods of automatic term recognition, in: National Center for Science Information Systems, 1996, pp. 1–22.

[3] F. Vezzani, G. M. Di Nunzio, V. Bonato, G. Silvello, Overview of the international workshop on definition and term extraction challenge (detech) 2026, in: Proceedings of the International Workshop on Definition and Term Extraction Challenge (DETECH) 2026, CEUR.org, Zadar, Croatia, 2026.

[4] M. T. Cabré, R. Estopà, J. Vivaldi, Automatic term detection: a review of current systems, in: Recent Advances in Computational Terminology, John Benjamins, Amsterdam, Philadelphia, 2001.
[5] M. Marciniak, P. Rychlik, A. Mykowiecka, Supporting terminology extraction with dependency parses, in: Proceedings of the 6th International Workshop on Computational Terminology, Marseille, France, 2020, pp. 72–79.
[6] G. Lame, Classement automatique de documents et analyse terminologique de corpus, in: Actes de la conférence TIA-2001, Nancy, France, 2001, pp. 149–158.
[7] B. Daille, Study and implementation of combined techniques for automatic extraction of terminology, in: Proceedings of the The Balancing Act : Combining Symbolic and Statistical Approaches to Language, Workshop at the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, NM, 1994, pp. 29–36.
[8] J. S. Justeson, S. M. Katz, Technical terminology: some linguistic properties and an algorithm for identification in text, Natural Language Engineering 1 (1995) 9–27.
[9] P. Drouin, Term extraction using non-technical corpora as a point of leverage, Terminology 9 (2003) 99–117.
[10] P. Velardi, M. Missikoff, R. Basili, Identification of relevant terms to support the construction of domain, in: Proceedings of the ACL-EACL Workshop on Human Language Technologies, Kluwer Academic Publisher, 2001.
[11] P. Drouin, Acquisition automatique des termes : l'utilisation des pivots lexicaux spécialisés, Ph.D. thesis, Université de Montréal, 2002.
[12] J. Dowdall, MichaelHess, N. Kahusk, K. Kaljurand, M. Koit, F. Rinaldi, KadriVider, Technical terminology as a critical resource, in: Proceedings of LREC'2002, 2002.
[13] K. T. Frantzi, S. Ananiadou, H. Mima, Automatic recognition of multi-word terms: the C-Value/NC-Value method, International Journal on Digital Libraries 3 (2000) 115–130.
[14] K. T. Frantzi, S. Ananiadou, J. Tsujii, Automatic term recognition using contextual clues, in: Proceedings of the Second Workshop on Multilingality in software Industry: The AI Contribution (MULSAIC'97), , 15th International Joint Conference on Artificial Intelligence, IJCAI'97, Nagoya, Japan, 1997, pp. 73–79.
[15] G. Bordea, P. Buitelaar, T. Polajnar, Domain-independent term extraction through domain modelling, in: Proceedings of the 10th International Conference on Terminology and Artificial Intelligence (TIA 2013), France, 2013.
[16] A. Šajatović, M. Buljan, J. Šnajder, B. Dalbelo Bašić, Evaluating automatic term extraction methods on individual documents, in: Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), Florence, Italy, 2019, pp. 149–154.
[17] E. Bolshakova, N. Loukachevitch, M. Nokel, Topic models can improve domain term extraction, in: P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, E. Yilmaz (Eds.), Advances in Information Retrieval, Berlin, Heidelberg, 2013, pp. 684–687.
[18] A. Rigouts Terryn, P. Drouin, V. Hoste, E. Lefever, Analysing the impact of supervised machine learning on automatic term extraction: HAMLET vs TermoStat, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), INCOMA Ltd., Varna, Bulgaria, 2019, pp. 1012–1021.
[19] E. Amjadian, D. Inkpen, T. Paribakht, F. Faez, Local-global vectors to improve unigram terminology extraction, in: Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016), Osaka, Japan, 2016, pp. 2–11.
[20] M. Kucza, J. Niehues, T. Zenkel, A. Waibel, S. Stüker, Term extraction via neural sequence labeling a comparative evaluation of strategies using recurrent neural networks, 2018, pp. 2072–2076.
[21] B. Q. Zadeh, S. Handschuh, Investigating context parameters in technology term recognition, in: Proceedings of the COLING Workshop on Synchronic and Diachronic Approaches to Analyzing Technical Language, Dublin, Ireland, 2014, pp. 1–10.
[22] S. Pollak, A. Repar, M. Martinc, V. Podpečan, Karst exploration: Extracting terms and definitions from karst domain corpus, 2020.
[23] A. Hazem, M. Bouhandi, F. Boudin, B. Daille, TermEval 2020: TALN-LS2N system for automatic

term extraction, in: Proceedings of the 6th International Workshop on Computational Terminology, Marseille, France, 2020, pp. 95–100.
[24] H. T. H. Tran, M. Martinc, J. Caporusso, J. Delaunay, A. Doucet, S. Pollak, Recent advances in automatic term extraction: A comprehensive survey, ACM Comput. Surv. 58 (2026).
[25] G. M. Di Nunzio, S. Marchesin, G. Silvello, A systematic review of automatic term extraction: What happened in 2022?, Digital Scholarship in the Humanities 38 (2023) i41–i47.
[26] H. Schmid, Probabilistic part-of-speech tagging using decision trees, in: Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, 1994.
[27] S. Aubin, T. Hamon, Improving term extraction with terminological resources, in: T. Salakoski, F. Ginter, S. Pyysalo, T. Pahikkala (Eds.), Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006), number 4139 in LNAI, Springer, 2006, pp. 380–387.
[28] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 7 (1997) 1735–1780.
[29] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, Psychological Review 65 (1958) 386–408.