University of Szeged
Research Group on Artificial Intelligence

Feature Engineering for Domain Independent Named Entity Recognition and Biomedical Text Mining Applications

Summary of the PhD Thesis

by

György Szarvas

Supervisor:

Prof. János Csirik

Szeged
2008

Introduction

The booklet summarizes the scientific results of the author of the PhD dissertation entitled *Feature Engineering for Domain Independent Named Entity Recognition and Biomedical Text Mining Applications*. The dissertation concentrates on two key topics in artificial intelligence (AI): machine learning (ML) and its application to Natural Language Processing tasks.

The amount of information stored in electronic document form is growing at an exponentially increasing rate. Searching for and processing information from textual sources is increasingly time-consuming in many areas (like medicine, research or business) and is becoming infeasible to perform without computer assistance. Thus today the need for intelligent text processing applications that can supersede or assist human information search in text documents is strong. Even though the performance of computers is not as good as the performance of humans in most complex information processing tasks, computers also have some obvious advantages to humans in their capacity of processing and the precision in performing well-defined tasks (e.g. indexing the whole World Wide Web).

An emerging field of Natural Language Processing is Text Mining, which seeks to automatically process large amounts of unstructured text; that is, to locate and classify relevant items of information and populate some sort of structured data collection for human processing. This task obviously requires at least a limited understanding of the text itself and the induction of new complex patterns that simulate human information search, which makes text mining tasks more complex and challenging than traditional keyword-lookup based information retrieval tasks.

Aim of the thesis

In the thesis we present several practical text mining applications developed together with colleagues, from a feature representation point of view. The applications themselves cover a wide range of different tasks from entity recognition (word sequence labelling) to multi-label document classification and different domains (from business news texts to medical records / biological scientific papers). Our aim is to demonstrate that task-specific feature engineering is beneficial to the overall performance and for specific text mining tasks, it is feasible to construct systems that are useful in practice and even compete human performance in processing the majority of textual data.

Structure of the thesis

This booklet is organized similarly to the thesis itself. The booklet is divided into two major parts, the first one focusing on Named Entity Recognition (thesis chapters 2-4) and the second on Text Classification (thesis chapters 5-7). At the end of the booklet we discuss the most important contributions of the author to the research and development described in the thesis, and we also list the author's contributions for each cited paper (the papers that discuss the same topics and results as the thesis).

Part I – Entity Recognition tasks

The identification and classification of rigid designators [1] like proper nouns, acronyms of names, time expressions, measures, IDs, e-mail addresses and phone numbers in plain text is of key importance in numerous natural language processing applications. The special characteristic of these rigid designators (as opposed to common words) is that they have no *meaning* in the traditional sense but they refer to one or more entities of the world uniquely (references). These text elements are called Named Entities (NEs) in the literature.

Named Entities generally carry important information about the text itself, and thus are targets for Text Mining [2]. Another task where NEs have to be identified is Machine Translation which has to handle proper nouns and common words in a different way due to the specific translation rules that apply to names [3]. Entity co-reference [4] and disambiguation [5] (two related tasks) is also an important task in Information Retrieval since a major part of queries are entity names that are highly ambiguous.

Because of the above, the very first step in almost every Text Mining task is to detect names in the text that belong to task-specific entity types. These tasks are the so-called Named Entity Recognition (NER) problems, where one tries to recognise (single or subsequent) tokens in text that together constitute a rigid designator phrase, and to determine the category type to which these phrases belong. Categorisation is always task specific, as different kinds of entities are important in different domains. Sometimes entity recognition itself can be a standalone application, as in the case of anonymisation issues, where no further processing is required when all the name phrases have been located in the text.

NER tasks can be solved using labelled corpora and statistical methods that induce NE tagging rules by discovering patterns in the manually annotated source text. Being a simple but crucial task, NER has been evaluated for various domains and languages. The variety of languages for which a major evaluation campaign include English [6], German [7], Dutch, Spanish [8], Chinese [9], Japanese [10] just to name a few, while domains where NER has been (and is) studied extensively includes the task of processing economic, sports and political news [7], medical texts [11], chemical [12], biological texts [13] or military documents [14]. In this study we deal with Hungarian NER and English newswire and medical NER.

Though the nature of the information that is important and is thus the target for recognition differs from application to application, these tasks can be handled with such models that are, in a limited sense, language [7] [15] (and domain [16] [17]) independent. The language and domain independence of NER systems means that a similar algorithm is capable of solving various tasks, independently of the target language and domain, as long as a labelled corpus for a particular language/domain pair is available and the entity types to be recognised are more or less similar. Cross language and/or cross domain recognition where systems are trained and used in different languages or domains has also been widely studied, but such scenarios are beyond the scope of this work, hence they will not be discussed here. We will focus on the recognition of proper names in multiple languages (Hungarian and English) and multiple domains (newswire and medical texts) and the recognition of a few other entity types like dates, IDs, etc. in English medical documents.

Example name tagging tasks

Now we will give a very brief introduction to some specific name tagging tasks in order to give the reader a better insight into the nature of NER tasks. The examples listed here include those problems that we will address later on in this thesis.

English newswire NER

We dealt with the NER evaluation task of the Computational Natural Language Learning (CoNLL) 2003 [7] conference for English language. Here the goal was the correct identification of personal,

organization and location names, along with other proper nouns treated as miscellaneous entities in texts of news press releases of Reuters Inc. from 1996.
An example of English NER is

- [U.N.]$_{ORGANISATION}$ official [Rolf Ekeus]$_{PERSON}$ heads for [Baghdad]$_{LOCATION}$.

Hungarian newswire NER

We addressed a NER task similar to the CoNLL 2003 guidelines for Hungarian language[18]. Thus we had to distinguish between person, organization and location names, and miscellaneous entities, and used texts from the Szeged TreeBank[19] of short press releases of the Magyar Távirati Iroda.
An example of Hungarian NER is

- A pénzügyi kockázatok kezeléséről kétnapos nemzetközi konferenciát tartanak csütörtökön és pénteken [Budapesten]$_{LOCATION}$ - mondta [Kondor Imre]$_{PERSON}$, a [Magyarországi Kockázatkezelők Egyesületének]$_{ORGANISATION}$ elnöke szerdán [Budapesten]$_{LOCATION}$ a sajtótájékoztatón.

English medical NER

For medical texts, an important use of NER is the automatic anonymisation of medical reports, to facilitate information exchange/access and respect individual patient rights (the protection of personal data). According to the guidelines of Health Information Portability and Accountability Act (HIPAA) of the US, the medical records released must be free of seventeen categories of textual Personal Health Information (PHI), out of which 8 actually appeared in the discharge summaries we used: first and last names of patients, their health proxies, and family members; the patient's age (if above 89 years old); doctors' first and last names; identification numbers; telephone, fax, and pager numbers; hospital names; geographic locations; and dates. To develop and test our model we used the de-identification dataset of the I2B2 Workshop on Challenges in Natural Language Processing for Clinical Data [11].
An example of the de-identification task is

- Mr. [Cornea]$_{PATIENT}$ underwent an ECHO and endoscopy at [Ingree and Ot of Weamanshy Medical Center]$_{HOSPITAL}$ on [April 28]$_{DATE}$.

The statistical NER system we developed

To solve the above problems we developed a statistical NER system that performed well across languages (English and Hungarian) and domains (newswire press releases and medical reports) with only slight modifications needed to port the system from the newswire domain to medical texts - we added a few features that exploit the specific characteristics of medical texts and a loop to the training process to achieve an even better performance. Our results showed that the model was successful even without these (fine tuning) domain extensions.

The NER system we developed treats the NER problem as the classification of separate tokens. Using labeled corpora of about 200000 tokens in size, we applied a decision tree classifier (C4.5 [20]) and boosting (AdaBoostM1 [21]) to NER, two algorithms that are well-known from the machine learning literature.

To solve a similar NER problem in different settings, we use the same learning model, and the same or very slightly modified feature set. Of course, most features that have an external source (lists, frequency information, etc.) are customized to the actual task by using a different source for calculating feature values, i.e. Hungarian NER uses Hungarian lists, English NER uses English lists, medical NER uses medical term lists, etc. Our general classifier model exploits features of several different types (a more detailed description is given in the corresponding thesis chapter), including:

- **gazetteers** of unambiguous NEs from the train data: we used the NE phrases which occur more than five times in the train texts and had the same label in over 90% of the cases,
- **dictionaries** of first names, company types, sport teams, denominators of locations (mountains, city) and so on: we collected special English lists from the Internet,
- **orthographical features**: capitalization, word length, common bit information about the word form (contains a digit or not, has an uppercase character inside the word, regular expressions and so on). We collected the most representative character level bi/trigrams from the train texts assigned to each NE class,
- **frequency information**: frequency of the token, the ratio of the token's capitalized and lowercase occurrences, the ratio of capitalized and sentence start frequencies of the token,
- **phrasal information**: chunk codes and forecasted class of a few preceding words (we carried out an online evaluation),
- **contextual information**: POS codes, sentence position, document zone (title or body), topic code, trigger words (the most frequent and unambiguous tokens in a window around the NEs) from the train text and whether the word is inside quote marks or not.

Owing to the beneficial characteristics of decision tree learning and the compact feature representation we developed (using fewer than 200 features for the general NER task, and omitting the word form itself from the classification process), our model is fast to train and evaluate, and performed well on standard evaluation datasets.
Our domain and language independent model achieved:

- an 89.02% F measure on the CoNLL 2003 evaluation set
- a 94.76% F measure on the Hungarian Named Entity Corpus
- a 94.34% F measure on the de-identification challenge of the I2B2 workshop.

Domain extensions improved the performance of our system on medical texts and our model gave an F measure of 97.64%. All these evaluations correspond to phrase-level equal-weighted F measures on each entity class.

Part II – Text Classification tasks in biomedical texts

The human processing of textual data (system logs, medical reports, newswire articles, customer feedback records, etc.) is a laborious and costly process, and is becoming unfeasible with the increasing amount of information stored in documents. There is a growing need for solutions that automate or facilitate the information processing workflow that is currently performed by humans. Thus today the automatic classification of free texts (either assertions or longer documents) based on their content and converting textual data to practical knowledge is an important subtask of Information Extraction.

Many text processing tasks can be formulated as a classification problem and solved effectively with Machine Learning methods [22] that are capable of uncovering the hidden structure in free text, assuming that labelled examples are on hand to train the automatic systems on. These solutions go one step beyond simple information retrieval (that is, providing the user with the appropriate documents using keyword lookup and relevance ranking), as they require the (deep or shallow) understanding of the text itself. The systems have to handle synonymy, transliterations and language phenomena like negation, sentiment, subjectivity and temporality [23].

A major application domain of practical language technology solutions is the field of Biology and Medicine [24]. Experts in these fields usually have to work with large collections of documents in everyday work in order to carry out efficient research (reading scientific papers, patents, or reports on earlier experiments in the subject) or decision making (reports on examination of former patients with similar symptoms or diseases).

Even though language or domain independent models would be desirable as well, solutions of such generality are in many cases beyond the scope of current state-of-the-art NLP technology. Economic aspects thus motivate the development of more specific solutions for unique concrete problems. In this thesis we will focus on the issues associated with biological and medical text processing.

Example text classification tasks

Here we will give a brief introduction to some specific tasks in order to give the reader a better insight into the nature of assertion or document level text classification. The examples listed here include those problems that will be addressed later on in this thesis.

Smoker status extraction from medical discharge summaries

The main purpose of processing medical discharge records is to facilitate medical research carried out by physicians by providing them with statistically relevant data for analysis. An example of such an analysis might be a comparison of the runoff and effects of certain diseases among patients with different social habits [25]. The evidence drawn from the direct connection between social characteristics and diseases (like the link between smoking status and lung cancer or asthma) is of key importance in treatment and prevention issues.

Such points can be deduced automatically by applying statistical methods on large corpora of medical records. Here we used the 'smoker status' dataset of the I2B2 Workshop on Challenges in Natural Language Processing for Clinical Data [26]. The task in this case is to classify the medical records into the following five semantic classes based on the smoking status of the patient being examined:

- non-smoker: the patient has no smoking history,
- current smoker: he/she is an active smoker,
- past smoker: the patient had not smoked for at least one year,

- smoker: when the document contains no information about his current or past smoker status, but he/she has smoking history,

- unknown: the report contains no information about the patient smoking status.

A sample assertion on patient smoking status in a discharge summary is:

- *The patient is a 60 yo right handed gentleman with a 20-years history of heavy smoking. Agreed to participate in a smoking cessation program. (current smoker)*

Detection of speculations in assertions

The highly accurate identification of several regularly occurring language phenomena like the speculative use of language [27] [28], negation and past tense (temporal resolution) is a prerequisite for the efficient processing of biomedical texts. In various Text Mining tasks, relevant statements appearing in a speculative context are treated as false positives. Hence hedge detection seeks to perform a kind of semantic filtering of texts; that is it tries to separate factual statements from speculative/uncertain ones. For biological scientific texts, we used a corpus consisting of articles on the fruit fly, provided by [29], and also used a small annotated corpus of 4 BMC Bioinformatics articles for external-source-evaluation. To evaluate our models in the medical domain, we used the standard dataset provided for the International Challenge on Classifying Clinical Free Text Using Natural Language Processing and a rule-based ICD-9 coder system constructed by us to provide false positive ICD-9 labels for automatic hedge dataset generation.
Two examples of speculative assertions in biological scientific texts are:

- *Thus, the D-mib wing phenotype may result from defective N inductive signaling at the D-V boundary.*

- *A similar role of Croquemort has not yet been tested, but seems likely since the crq mutant used in this study (crqKG01679) is lethal in pupae.*

Two examples of speculative assertions radiology reports are:

- *Findings suggesting viral or reactive airway disease with right lower lobe atelectasis or pneumonia.*

- *Right middle lobe infiltrate and/or atelectasis.*

Automatic ICD-9-CM coding of radiology reports

The assignment of International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) codes serves as a justification for carrying out a certain procedure. This means that the reimbursement process by insurance companies is based on the labels that are assigned to each report after the patient's clinical treatment. The approximate cost of ICD-9-CM coding clinical records and correcting related errors is estimated to be about $25 billion per year in the US [30].

Since the ICD-9-CM codes are mainly used for billing purposes, the task itself is commercially relevant: false negatives (i.e. overlooked codes that should have been coded) will cause a loss of revenue to the health institute, while false positives (the over coding of documents) is penalised by a sum three times higher than that earned with the superfluous code, and also entails the risk of prosecution to the health institute for fraud. Thus there is a desperate need for high-performance automatic ICD-9 coding systems. Here we used the standard dataset provided for the shared task International Challenge on Classifying Clinical Free Text Using Natural Language Processing on ICD-9-CM coding of radiology reports [31].
Two examples of ICD-9-CM coding of radiology reports are:

- *CODES: 486; 511.9*
  HISTORY: Right lower lobe pneumonia, cough, followup.
  IMPRESSION: Persistent right lower lobe opacity with pleural effusion present, slightly improved since prior radiograph of two days previous.

- *CODES: 593.89; V13.02*
  HISTORY: 14-year - old male with history of a single afebrile urinary tract infection in January with gross hematuria for a week. The patient was treated with antibiotics.
  IMPRESSION: Mild left pyelectasis and ureterectasis. Otherwise normal renal ultrasound. The bladder appears normal although there is a small to moderate post void residual.

The statistical assertion/document classifier systems we developed

To solve the above tasks and get a satisfactory performance we developed machine learning models for each that differ slightly from each other. The cost of building corpora of reasonable size for these tasks is very high; and thus to have labeled corpus of a similar size to that for entity tagging problems was not deemed feasible. Instead, we had to develop solutions that can infer the structure and knowledge hidden in the text from a few hundred (or few thousand) examples. Another option is to gather labeled examples fully or semi-automatically (within the scope of weakly supervised learning), but in such cases significant noise in the semi-automatic labeling has to be dealt with. Despite the nice results we obtained on these tasks the portability of the learned hypotheses suffered from this lack of training data, as our experiments revealed.

Smoker status extraction from medical discharge summaries

To classify smoker status in discharge summaries, we applied a sentence-level classifier based on the VSM representation of discharge summaries. We extended a unigram representation with complex features of pre-classified bigrams and trigrams (based on their meaning) and simple syntactic features plus negation detection as well. Our experiments showed that the features we introduced for smoker status classification were more helpful for learning than a simple VSM used by earlier approaches (feature selection methods chose our complex features more often than unigrams). The final classification of our system was based on the majority voting of several classifiers (C4.5 decision tree, Multi-Layer Perceptron and Support Vector Classifier). The solution we proposed for smoker status detection proved to be particularly efficient in discarding irrelevant documents (unknown class) and non-smokers, thus it could be used as a pre-processor for human processing/annotation. Our model achieved an overall accuracy of 86.54% on the I2B2 challenge evaluation set, close to the best solution that was entered in the challenge.

Detection of speculations in assertions

Here we used weakly supervised settings for biological texts and no supervision for medical data to acquire training sets for detecting hedges in biological or medical texts (at the sentence level). To classify sentences into speculative and non-speculative assertions we applied a Maximum Entropy classifier [32] and vector space model (VSM) to represent the examples. Uni-, bi- and trigrams of words were all included in the VSM representation. After a careful selection of relevant hedge keywords that involved a ranking and filtering of keywords via a modified class-conditional probability score (only the best 2 keywords received credit for appearing in speculative sentences) and then sorting the best candidates according to the $P(spec)$ scores given by the Maximum Entropy classifier, we managed to filter out the best speculative keywords from a training set constructed by using minimal supervision (biological scientific texts) or no supervision at all (radiology reports). This procedure resulted in an

$F_{\beta=1}(spec)$ score of 85.08% for biological scientific papers, and an $F_{\beta=1}(spec)$ score of 82.07% for clinical free-texts.

Having selected the most relevant keywords, our hedge classifier simplified to a simple keyword matching routine that predicted speculative label every time a strong keyword was present. We suggest the use of 2 or 3 word long phrases to capture the rare non-speculative uses of otherwise strong keywords as a possible solution for further improving the model. This idea would require a labeled corpora of reasonable size to implement and evaluate, though.

The automatic ICD-9-CM coding of radiology reports

For the automated clinical coding of medical free texts we applied a hybrid rule-based and statistical model. A unique feature of the task was that expert coding guides that describe the principles of ICD-9-CM coding for humans were on hand and these guides were good sources for the swift implementation of an ICD-9-CM coding expert system. Thus in our study we focused on the possible ways of exploiting labeled data to fine-tune such an expert rule-based system.

First we developed a simple rule-based system for ICD-9 coding using one of the several online coding guides available. Then we trained statistical models using the codes assigned by the rule-based system to model the inter-label dependencies between disease and associated symptom codes. To address the second major deficiency of rule-based systems based on coding guides – i.e. terminology missing from the lists in the coding guide (rare transliterations and abbreviations, etc.) – we also applied a statistical approach. In this case we trained classifiers for the false negative codes of the initial rule-based system.

With this hybrid rule-based and statistical model we got a very good performance, one very close to an entirely manually constructed system with a moderate development cost that would make the implementation of our system feasible for a large set of codes as well (while developing a hand-crafted system for several hundreds or thousands of codes would be problemmatic indeed).

Summary

Summary by chapters

Here we summarise our findings for each chapter of the thesis and provide the relation of each paper referred to in the thesis and the results described in different chapters in a table.

The thesis is divided into two major parts, one dealing with Named Entity Recognition problems and another that focuses on Text Classification tasks. Here we list our most important findings for each chapter.

- NER Chapters

  1. Hungarian NER Chapter

     For NER in Hungarian the author participated in the creation of the first Hungarian NER reference corpus which allowed researchers to investigate statistical approaches to Entity Recognition in Hungarian texts. This is a joint, inseparable contribution between the authors of [18] and the linguist colleagues who carried out the annotation work of the corpus.

     Together with his colleagues, the author designed a suitable feature representation for training machine learning models and set up an efficient learning model on the corpus that achieved a phrase level F measure performance of 94.76%. In the construction of the Named Entity Recognition system, the author made major contributions in designing the feature representation for learning algorithms.

     These results are described in [18], [33] and [34].

  2. English NER Chapter

     The author participated in adapting a NER system designed for the Hungarian language to a similar task in English. Together with his colleagues, the author extended the feature representation for training machine learning models and used the same, efficient learning model that was introduced for Hungarian NER. This system attained a phrase level F measure score of 89.02%.

     The author also participated in the development of a MaxEnt-based system for the metonymy resolution shared task of SemEval-2007 [35]. In the NE-metonymy classifier that was submitted to the challenge by the author and his colleagues, the author is responsible for the web-based approach that was designed to remove inflectional affixes from Named Entities and was used successfully as a feature to classify org-for-product metonymies.

     When constructing the English Named Entity Recognition system, the author made major contributions in designing the feature extensions for learning algorithms.

     The author investigated corpus frequency based heuristics that were capable of fine tuning NER systems by eliminating certain typical errors of NER systems. These heuristics were then altered to provide a heuristic solution to Named Entity lemmatisation, a problem that arises both in English (plural and possessive markers) and in Hungarian (agglutinative characteristic of the language). The author and his collegues showed that corpus statistics can be utilised to solve NE lemmatisation with good accuracy. The author's contribution is the idea and general concept of using web frequency counts for Named Entity lemmatisation (NE normalisation or affixes as features for other tasks).

     These results are described in [34], [36], [37] and partly in [38].

  3. Anonymisation of Medical Records

     Together with his collegues, the author participated in the 2006 I2B2 shared task challenge on medical record de-identification. The major parts of the adaptation of the pre-existing

NER system, and the results achieved as a whole are the joint contribution of the co-authors. As our results show, the system we built via the domain adaptation of our newswire NER model is competitive with other approaches, which means that our architecture is capable of solving NER tasks language and domain independently, with minimal adaptation effort.

In particular, the author made major contributions to the customisation of the feature representation, i.e. the development of novel features specifically for the medical domain. These novel features were helpful in achieving a state-of-the-art performance (our model had the best phrase-level 8-way F measure and second best token-level 9-way accuracy).

These results are described in [39].

- Text Classification Chapters

  1. Smoker status classification in discharge summaries

     Together with his collegues, the author participated in the 2006 I2B2 shared task challenge on patient smoking status classification from medical records. The system and the overall results we submitted are a shared and indivisible contribution of the co-authors.

     In particular, the author made major contributions to the design of the feature representation, i.e. the development of features used by previous studies and novel ones specifically for the medical domain which tried to group more or less similar examples together by exploiting the syntactic or semantic classification of phrases. The main reasoning for having these novel features was to reduce the effects of a small sample size. These novel features were helpful in achieving a good performance (they appeared among the top ranked attributes using 2 different feature selection methods).

     These results are described in [40].

  2. Hedge Classification in biomedical texts

     All the contributions in the corresponding chapter are independent results of the author. The major findings of this thesis are the construction of a complex feature ranking and selection procedure that successfully reduces the number of keyword candidates (those having the highest class-conditional probability for hedge class) without excluding helpful hedge keywords.

     We also demonstrated that with a very limited amount of expert supervision in finalising the feature representation, it is possible to build accurate hedge classifiers from semi-automatically or automatically collected training data.

     We extended the scope of evaluations to two applications with different kinds of texts involved (scientific articles used in previous works, and also medical free texts).

     We extended the feature representation used by previous approaches to 2-3 word-long phrases and an evaluation of the importance of longer keywords in hedge classification.

     We demonstrated (using a small test corpora of biomedical scientific papers from a different source) that hedge keywords are highly task-specific and thus constructing models that generalise well from one task to another is not feasible without a noticeable loss in accuracy.

     These results are described in [41] and partly in [42].

  3. ICD-9-CM coding in radiology reports

     Together with his collegues, the author participated in the 2007 CMC shared task challenge on automated ICD-9-CM coding of medical free texts using Natural Language Processing. The major steps of the development of the system as a whole that was submitted to the challenge, and the results achieved are a shared and indivisible contribution of the co-authors.

     In particular, the author made a major contribution to the development of a basic and an entirely hand-crafted rule-based classifier; the design, implementation and interpretation of the complex inter-annotator agreement analysis and the design of the machine learning model for discovering inter-label dependencies from the labeled corpus.

These results are described in [43].

	HunNER	EngNER	DE-ID	SMOKER	HEDGE	ICD-9
LREC[18]	●					
ACTA[33]	●					
DS2006[34]	●	●				
SEMEVAL[38]		●				
ICDM2007[36]		●				
TSD2008[37]		●				
JAMIA[39]			●			
WSEAS[40]				●		
ACL[41]					●	
BIONLP[42]					●	
LBM2007[43]						●

Table 1: The relation between the thesis topics and the corresponding publications.

Summary by papers

Here we list the most important results in each paper that are regarded as the author's own contributions. We mention here that system performance scores (i.e. the overall results) are always counted as a shared contribution and not listed here, as several authors participated in the development of the systems described in the cited papers. The only exception is [41], which describes only the author's own results. [18] has been omitted from the list as all the results described in this paper are counted as shared contributions of the authors. For [38] the author made only marginal contributions.

- ACTA[33]
  - The construction of a feature representation for Hungarian NER.
  - Compact representation.
  - Frequency features.
- DS2006[34]
  - Description of feature space extensions for English NER.
- SEMEVAL[38]
  - The plural feature for NE-metonymy resolution.
- ICDM2007[36]
  - Using web frequency data for identifying consecutive NEs.
- TSD2008[37]
  - The idea and general concept of using web frequency counts for Named Entity lemmatisation (NE normalisation or affixes as features for other tasks)

- JAMIA[39]

  – The extension of the feature space with respect to the chief characteristics of medical texts.

  – The iterative learning/feature generation approach.

- WSEAS[40]

  – The use of token bi- and trigram features.

  – The use of deep knowledge features (pre-classified bigrams, syntactic information, negation).

  – The execution of feature selection methods for getting a suitable set of features for smoker classification.

- ACL[41]

  – All of the results in the paper.

- BIONLP[42]

  – Some of the general principles of negation, hedging and their scope annotation.

- LBM2007[43]

  – Detailed performance and annotator agreement analysis.

  – Complex features for discovering label-dependecies with machine leraning models.

  – The construction of a basic rule-based system that served as the basis for further developments, and an entirely hand-crafted system for comparison.

References

[1] Kripke S: *Naming and Necessity*. Harvard University Press 1972.

[2] Jurafsky D, Martin JH: *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall, second edition 2008, [http://www.amazon.de/exec/obidos/redirect?tag=citeulike01-21\&amp;path=ASIN/013122798X].

[3] Babych B, Hartley A: **Improving Machine Translation Quality with Automatic Named Entity Recognition**. In *Proceedings of the 7th International EAMT workshop at EACL-2003*, Budapest, Hungary: Association for Computational Linguistics 2003:18–25.

[4] Nicolae C, Nicolae G: **BESTCUT: A Graph Algorithm for Coreference Resolution**. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia: Association for Computational Linguistics 2006:275–283, [http://www.aclweb.org/anthology/W/W06/W06-1633].

[5] Cucerzan S: **Large-Scale Named Entity Disambiguation Based on Wikipedia Data**. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* 2007:708–716.

[6] Chinchor NA: **Overview of MUC-7/MET-2.** In *Proceedings of the Seventh Message Understanding Conference (MUC-7)* 1998.

[7] Tjong Kim Sang EF, De Meulder F: **Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition.** In *Proceedings of CoNLL-2003*. Edited by Daelemans W, Osborne M, Edmonton, Canada 2003:142–147.

[8] Tjong Kim Sang EF: **Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition**. In *Proceedings of CoNLL-2002*, Taipei, Taiwan 2002:155–158.

[9] Tou Ng H, Kwong OOY (Eds): *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, Sydney, Australia: Association for Computational Linguistics 2006.

[10] Sekine S, Isahara H: **IREX: IR and IE evaluation project in Japanese** 2000, [citeseer.ist.psu.edu/sekine00irex.html].

[11] Uzuner O, Luo Y, Szolovits P: **Evaluating the State-of-the-Art in Automatic De-identification**. *J Am Med Inform Assoc* 2007, **14**(5):550–563, [http://www.jamia.org/cgi/content/abstract/14/5/550].

[12] Corbett P, Batchelor C, Teufel S: **Annotation of Chemical Named Entities**. In *Biological, translational, and clinical language processing*, Prague, Czech Republic: Association for Computational Linguistics 2007[http://www.aclweb.org/anthology/W/W07/W07-1008].

[13] Kim J, Ohta T, Tsuruoka Y, Tateisi Y, Collier N: **Introduction to the bio-entity recognition task at JNLPBA**. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA), Geneva, Switzerland*. Edited by Collier N, Ruch P, Nazarenko A 2004:70–75.

[14] Grishman R, Sundheim B: **Message Understanding Conference-6: a brief history**. In *Proceedings of the 16th conference on Computational linguistics*, Morristown, NJ, USA: Association for Computational Linguistics 1996:466–471.

[15] Cucerzan S, Yarowsky D: **Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence**. In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, College Park, MD, USA: Association for Computational Linguistics 1999:90–99.

[16] Kozareva Z: **Bootstrapping Named Entity Recognition with Automatically Generated Gazetteer Lists**. In *Proceedings of the Student Research Workshop at 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy: Association for Computational Linguistics 2006:15–21.

[17] Lee HS, Park SJ, Jang H, Lim J, Park SH: **Domain Independent Named Entity Recognition from Biological Literature**. In *Proceedings of The 15th International Conference on Genome Informatics*, Yokohama, Japan 2004.

[18] Szarvas Gy, Farkas R, Felföldi L, Kocsor A, Csirik J: **A highly accurate Named Entity corpus for Hungarian**. In *Proceedings of Language Resources and Evaluation Conference* 2006.

[19] Csendes D, Csirik J, Gyimóthy T, Kocsor A: **The Szeged Treebank**. In *TSD* 2005:123–131.

[20] Quinlan JR: *C4.5: Programs for Machine Learning*. Morgan Kaufmann 1993.

[21] Schapire R: **The boosting approach to machine learning: An overview**. In *Proceedings of MSRI Workshop on Nonlinear Estimation and Classification*, Berkeley, CA, USA 2001.

[22] Sebastiani F: **Machine learning in automated text categorization**. *ACM Comput. Surv.* 2002, **34**:1–47, [http://portal.acm.org/citation.cfm?id=505282.505283].

[23] Shanahan JG, Qu Y, Wiebe J: *Computing Attitude and Affect in Text: Theory and Applications (The Information Retrieval Series)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc. 2005.

[24] Ananiadou S, Mcnaught J: *Text Mining for Biology And Biomedicine*. Norwood, MA, USA: Artech House, Inc. 2005.

[25] Zeng Q, Goryachev S, Weiss S, Sordo M, Murphy S, Lazarus R: **Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system**. *BMC Medical Informatics and Decision Making* 2006, **6**:30, [http://www.biomedcentral.com/1472-6947/6/30].

[26] Uzuner O, Goldstein I, Luo Y, Kohane I: **Identifying Patient Smoking Status from Medical Discharge Records**. *J Am Med Inform Assoc* 2008, **15**:14–24, [http://www.jamia.org/cgi/content/abstract/15/1/14].

[27] Light M, Qiu XY, Srinivasan P: **The Language of Bioscience: Facts, Speculations, and Statements In Between**. In *HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases*. Edited by Hirschman L, Pustejovsky J, Boston, Massachusetts, USA: Association for Computational Linguistics 2004:17–24.

[28] Hyland K: **Hedging in Academic Writing and EAP Textbooks**. *English for Specific Purposes* 1994, **13**(3):239–256.

[29] Medlock B, Briscoe T: **Weakly Supervised Learning for Hedge Classification in Scientific Literature**. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic: Association for Computational Linguistics 2007:992–999, [http://www.aclweb.org/anthology/P/P07/P07-1125].

[30] Lang D: **Consultant Report - Natural Language Processing in the Health Care Industry**. *PhD thesis*, Cincinnati Children's Hospital Medical Center 2007.

[31] Pestian JP, Brew C, Matykiewicz P, Hovermale D, Johnson N, Cohen KB, Duch W: **A shared task involving multi-label classification of clinical free text**. In *Biological, translational, and clinical language processing*, Prague, Czech Republic: Association for Computational Linguistics 2007:97–104, [http://www.aclweb.org/anthology/W/W07/W07-1013].

[32] Berger AL, Pietra SD, Pietra VJD: **A Maximum Entropy Approach to Natural Language Processing**. *Computational Linguistics* 1996, **22**:39–71, [citeseer.ist.psu.edu/berger96maximum.html].

[33] Farkas R, Szarvas Gy, Kocsor A: **Named entity recognition for Hungarian using various machine learning algorithms**. *Acta Cybern.* 2006, **17**(3):633–646.

[34] Szarvas Gy, Farkas R, Kocsor A: **A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms**. In *Discovery Science* 2006:267–278.

[35] Markert K, Nissim M: **SemEval-2007 Task 08: Metonymy Resolution at SemEval-2007**. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic: Association for Computational Linguistics 2007:36–41, [http://www.aclweb.org/anthology/W/W07/W07-2007].

[36] Farkas R, Szarvas Gy, Ormándi R: **Improving a State-of-the-Art Named Entity Recognition System Using the World Wide Web**. In *Industrial Conference on Data Mining* 2007:163–172.

[37] Farkas R, Vincze V, Nagy I, Ormándi R, Szarvas Gy, Almási A: **Web based lemmatisation of Named Entities**. In *Accepted for 11th International Conference on Text, Speech and Dialogue* 2008.

[38] Farkas R, Simon E, Szarvas Gy, Varga D: **GYDER: Maxent Metonymy Resolution**. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic: Association for Computational Linguistics 2007:161–164, [http://www.aclweb.org/anthology/W/W07/W07-2033].

[39] Szarvas Gy, Farkas R, Busa-Fekete R: **State-of-the-art anonymisation of medical records using an iterative machine learning framework**. *J Am Med Inform Assoc* 2007, **14**(5):574–580, [http://www.jamia.org/cgi/content/abstract/M2441v1].

[40] Szarvas Gy, Iván S, Bánhalmi A, Csirik J: **Automatic Extraction of Semantic Content from Medical Discharge Records**. *WSEAS Transaction on Systems and Control* 2006, **1**(2):312–317.

[41] Szarvas Gy: **Hedge classification in biomedical texts with a weakly supervised selection of keywords**. In *Accepted for the 45th Annual Meeting of the Association of Computational Linguistics*, Columbus, Ohio, United States of America: Association for Computational Linguistics 2008.

[42] Szarvas Gy, Vincze V, Farkas R, Csirik J: **The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts**. In *Accepted for Biological, translational, and clinical language processing (BioNLP Workshop of ACL)*, Columbus, Ohio, United States of America: Association for Computational Linguistics 2008.

[43] Farkas R, Szarvas Gy: **Automatic construction of rule-based ICD-9-CM coding systems**. *BMC Bioinformatics* 2008, **9**(3), [http://www.biomedcentral.com/1471-2105/9/S3/S10].