Research Group on Artificial Intelligence
of the Hungarian Academy of Sciences
and the University of Szeged

Machine Learning techniques for applied Information Extraction

Summary of the PhD Dissertation
by

Richárd Farkas

Advisor:
Prof. Dr. János Csirik

University of Szeged, Faculty of Science and Informatics
Doctoral School of Computer Science

2009

Introduction

The booklet summarizes the scientific results of the author's PhD dissertation entitled *Machine Learning techniques for applied Information Extraction*. The dissertation is concerned with two key topics in artificial intelligence, namely Machine Learning and its application to Natural Language Processing tasks.

Due to the rapid growth of the Internet and the globalisation process, the amount of available information is growing at an incredible rate. The greater part of the new data sources is in textual form (e.g. every web page contains textual information) that is intended for human readers, written in a local natural language. This amount of information requires the involvement of the computer in the processing tasks. The automatic or semi-automatic processing of raw texts requires special techniques. *Natural Language Processing* (NLP) is the field which deals with the understanding and generation of natural languages (i.e. languages written by humans) by the computer.

Information Extraction (IE) is a subfield of NLP whose goal is to automatically extract structured information from unstructured, textual documents like Web pages, corporate memos, news articles, research reports, e-mails, blogs and so on. The output is structured information which is categorized and semantically well-defined data, usually in a form of a relational database. Example IE applications include the gathering of data about companies, corporate mergers from the Web or protein-protein interactions from biological publications. For example, from the sentence "*Eric Schmidt joined Google as chairman and chief executive officer in 2001.*" we can extract the information tuple:

{COMPANY=*Google*, CEO=*Eric Schmidt*, CEO_START_DATE=*2001*}

The first approaches for Information Extraction tasks were based on hand-crafted expert rules. The construction and maintenance of such a rule system is very expensive and requires a decision system specialist. *Machine Learning* techniques construct decision rules automatically based on a training examples – a manually annotated corpus in Information Extraction. The cost of the training set's construction is less than the cost of a hand-written rule set because the former one requires just domain knowledge i.e. labelling examples instead of decision system engineering. This thesis is concerned with the investigation of Machine Learning tools for Information Extraction tasks and their application in particular domains.

Aim of the thesis

The chief aim of the PhD thesis was to examine various Machine Learning methods and discuss their suitability in real-world Information Extraction tasks. Among the Machine Learning tools, several less frequently used ones and novel ideas will be experimentally investigated and discussed. The problems themselves cover a wide range of tasks from language-independent and multi-domain Named Entity Recognition (word sequence labelling) to Name Normalisation and Opinion Mining.

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 supervised learning for Information Extraction (thesis chapters 3-5) while the second deals with the exploitation potential of external knowledge in Information Extraction tasks (thesis chapters 6-8). At the end of the booklet we discuss the most important contributions of the author to the results and methods presented in the thesis, and we also list the author's contributions for the more important cited publications of the author.

Part I – Supervised learning for Information Extraction tasks

The standard approaches for solving IE tasks work in a supervised setting where the aim is to build a Machine Learning model on training instances for forecasting on previously unseen instances.

Supervised learning approaches

When attempting to solve classification problems effectively, it is worth applying various types of classification methods as the *No Free Lunch Theorem* [1] states that there is no single learning algorithm which in any given task always achieves the best results.
The two main approaches for Machine Learning-based IE are:

Token-level models carries out classification where the aim is to assign the correct tag (label) for each token "independently" in a plain text. These models are capable of taking into account the relationships between consecutive words as well, collecting the relevant features into a window of appropriate size. This means that the dependence is incorporated in the feature set (e.g. it can contain a feature like *the previous token is listed in a gazetteer*). This type of modelling provides the opportunity of sampling from individual tokens. This is necessary for creating a balanced training data, which is sometimes beneficial to learning algorithms.

Sequential models regard the whole sentence as an instance, i.e. their output for one prediction is a sequence of labels for the tokens of the sentence. They ignore the assumption about the independence of the elements of the sequence. To handle inter-token dependencies a more complex model is required, so sequential models have a worse training time complexity compared to token-level ones.

We describe comparative experiments on these approaches using several learning algorithms on the *Named Entity Recognition* [2; 3] and *Metonymy Resolution* [4] tasks.

A *Named Entity* (NE) is a phrase in the text which uniquely refers to an entity in the world. It includes proper nouns, dates, identification numbers, phone numbers, e-mail addresses and so on. As the identification of dates and other simpler categories are usually carried out by hand-written regular expressions we will focus on proper names like organisations, persons, locations, genes or proteins.

The identification and classification of proper nouns in plain text is of key importance in numerous natural language processing applications. It is the first step of an IE system as proper names generally carry important information about the text itself, and thus are targets for extraction. Moreover *Named Entity Recognition* (NER) can be a stand-alone application as well [5] and besides IE, Machine Translation also has to handle proper nouns and other sort of words in a different way due to the specific translation rules that apply to them.

In linguistics metonymy means using one term, or one specific sense of a term, to refer to another, related term or sense. The metonymic usage of NEs is frequent in natural language. For example in the following example *Vietnam*, the name of a location, refers to an event (the war) that happened there [6]:

> *Sex, drugs, and Vietnam have haunted Bill Clinton's campaign.*

Metonymy Resolution attempts to make the automatic distinction among the metonymic senses of the NEs.

Meta-learning

Meta-learning is the combination of several learning models (which can be learnt by the same algorithm or by different ones). There are several well known meta-learning algorithms in the literature that can lead to a 'better' model than those serving as a basis for it. The success of hybrid combination approaches lies in tackling the problem from several aspects, so algorithms with inherently different theoretical bases are good subjects for voting and for other combination schemes.

We introduced several experimental results on Hungarian and English NER tasks [7; 8] got by several meta-learning schemes. The combination schemes applied achieved a significant improvement compared to the best base-learner and the results confirm that even the base-learners with lower accuracy have added value in a combination scheme.

Besides the experiments with known meta-learning algorithms, we introduced a novel meta-learning scheme [3]. This procedure exploits our rich feature set built for the NER tasks. Here, several overlapping smaller feature sets were selected from the whole set. These sets describe the tokens from different perspectives and an accurate model can still be built on them. Then learning models on each smaller feature set were applied and their predictions in a Stacking approach made the final decision. This procedure can be regarded as a bagging method where instead of training models on different entity sets the bags contain the same dataset, but from a different view as different features describe it.

We built a complex statistical NER system which employs several meta-learning methods [3]. Firstly, the full feature set – which contains mostly nominal attributes – is extracted from the raw text and split into the five (overlapping) chosen subsets. AdaBoosted C4.5 decision trees are trained on these feature subsets and the Stacking combination of their forecasts is the output of our "individual" system. Lastly, this forecast is post-processed and the text is labeled. Our NER system is competitive with the published state-of-the-art systems while it has a different theoretical background, which makes it an excellent candidate for a combination scheme with outer NER systems.

Adaptability of IE systems

Supervised systems usually predicate well on unseen instances if they share the characteristics of the training dataset. Hence, when the target texts are varying new training datasets are required. We discussed several situations where the datasets are varying for a particular task but the same learning procedure with minor modifications can be applied.

We carried out experiments on Hungarian and English NER datasets. The experimental results show that our NER systems applied on them has the same characteristics, its components have similar added values and can achieve good accuracies on both languages. These results are quite satisfactory if we take into account the fact that the results for English are by far the best known, while NLP tasks in Hungarian are usually more difficult to handle because Hungarian has many special (and from a statistical learning point of view, undesirable) characteristics. Our NER system remains portable across languages as long as language specific resources are available; and it can be applied successfully to languages with very different characteristics.

In 2007, we adapted our newswire NER model to the medical anonymisation task where the goal is to identify and classify Personal Health Information (like names of patients; doctors' names; identification numbers; telephone, fax, and pager numbers; hospital names; geographic locations; and dates) in discharge summaries. Three minor modifications of the system yielded a model which achieved top results on an international open shared task [9].

Later on, we used our English and Hungarian NER system as a submodule in a sentence alignment tool which was tested on several genre of texts [10], thus we obtained an insight into the performance of our NER systems on general texts where a domain-specific training set was not present. Sentence alignment establishes relations between sentences of a bilingual parallel corpus. This relation may not have just a one-to-one correspondence between sentences; there could be a many-to-zero alignment (in the case of insertion or deletion), many-to-one alignment or even many-to-many alignments. In

general, words which are written with a capital letter is not a good anchor. Hence we suggest modifying the base cost of a sentence alignment with the help of NER instead of a bilingual dictionary of anchor words or the number of capital letters in the sentences. This leads to a text-genre independent anchor method that does not require any anchor filtering at all.

As a preliminary step towards automatic domain adaptation, we statistically investigated the differences among domains, focusing on the IE tasks of negation and uncertainty assertions. These tasks are essential in most IE applications where, in general, the intention is to derive factual knowledge from textual data. Take, for example, the clinical coding of medical reports, where the coding of a negative or uncertain disease diagnosis may result in an over-coding financial penalty. To aid the development of uncertainty and negation detection systems we built the BioScope corpus [11]. The corpus consists of three parts, namely medical free texts, biological full papers and biological scientific abstracts; and it contains annotations at the token-level for negative and speculative keywords and at the sentence-level for their linguistic scope. The statistical figures of the subcorpora tell us that the same tasks on slightly different domains (clinical records, full biological papers and biological abstracts) or very similar tasks on the same domain (hedge and negation detection) can have quite different characteristics, thus the adaptation procedure here is not straightforward.

Part II – Exploitation of external knowledge in Information Extraction tasks

As we discussed in Part I, supervised methods require a training corpus – with an appropriate size – for every task (the difference among tasks can be just marginal). In Part II of the thesis, we investigate several approaches to exploit knowledge outside the training data, thus decreasing the required amount of training data to a minimum level.

Using the WWW as unlabeled corpus

The most widely used external resources in a Machine Learning task are unlabeled instances. The way of training a model by using unlabeled data, together with labeled data is called *semi-supervised learning*. Here the goal is to utilise the unlabeled data during the training on labeled ones.

The special nature of NLP problems requires special semi-supervised techniques. There are two key points among these special characteristics. First, complex statistics can be simply gathered from unlabeled texts owing to the sequential structure of languages. Such statistics can be word and character bi-, trigrams, token or phrase frequencies and models of language in a wider sense (not just the usual $P(w_t|w_{t-1})$ distribution). This kind of information can be incorporated into the feature space for each Machine Learning process.

Another unique characteristic of NLP applications is that they can utilise the World Wide Web (WWW). The WWW can be viewed as an almost limitless collection of unlabeled data. Moreover it can bring some dynamism to applications, as online data changes and rapidly expands with time, a system can remain up-to-date and extend its knowledge without the need for fine tuning, or any human intervention. We introduced our NER refinement and lemmatisation approaches which exploits the largest unlabeled corpus of the world, the WWW.

During an analysis of the errors made by our NER system, we discovered that a significant proportion of errors came from the machine's lack of access to human common knowledge. If it possed such knowledge the system could not make errors like tagging the phrase *'In New York'* or give a *location* label to *'Real Madrid'*. We introduced WWW-based post processing techniques in order to refine the labeling of our NER model:

- we looked for the most frequent roles in which the NEs are used for and overwrite the original class of the NE in special cases,
- we extended the boundaries of an NE phrase if the such a decision was made based on the WWW-frequencies of the original phrase and the possibly extended form,
- we separated long NE phrases if they are thought to be distinct, consecutive NEs based on WWW-frequencies.

After recognising NEs in texts, finding their lemmas (and inflectional affixes) can be useful for several reasons: the proper name can be stored in a normalised form (e.g. for indexing) and it may prove to be easier to classify a proper name in the corresponding NE category using its lemma than the affixed form. The lemmatisation of common nouns can be made simply by relying on a good dictionary of lemmas [12]. The problem of proper name lemmatisation is more complicated since NEs cannot be listed exhaustively, unlike common nouns, due to their diversity and increasing number. Lots of suffixes can be added to the noun phrases in Hungarian (e.g. *Invitelben*, where *Invitel* is the lemma and *-ben* means *'in'* or *Pannon*, with *-on* meaning *'on'*)), and they can bear the plural or genitive marker *-s* or *'s* in English (e.g. *Toyotas*). What is more, there are NEs that end in an apparent suffix (such as *Adidas*, *McDonald's* or *Philips*), but this pseudo-suffix belongs to the lemma of the NE and should not to be removed.

In order to be able to select the appropriate lemma for each problematic NE, we applied the following strategy. In step-by-step fashion, each ending that seems to be a possible suffix is cut off

the NE. Our key hypothesis is that the frequency of the lemma-candidates on the WWW is high – or at least the ratio of the full form and lemma-candidate frequencies is relatively high – with an appropriate lemma and low in incorrect cases.

Integrating expert systems into Machine Learning models

Apart from unlabeled texts, existing expert decision systems, manually built taxonomies or written descriptions can hold useful information about the Information Extraction task in question. A nice example for this is clinical IE where the knowledge of thousands years has been gathered into medical lexicons. On the other hand hospitals and clinics usually store a considerable amount of information (patient data) as free text, hence NLP systems have a great potential in aiding clinical research due to their capability to process large document repositories both cost and time efficiently. We introduced several ways of integrating the medical lexical knowledge into Machine Learning models – which were trained on free-text corpora – through two clinical IE applications, namely *ICD coding* and *obesity detection*.

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 [13]. The coding guidelines define the codes for each disease and symptom and also place restrictions on how and when certain codes can be applied.

We built an automated ICD coder and participated in the CMC shared task [9]. The system has several components; negation and assertion detection, term identification and multi-label classification. We presented three different methods for term identification (the key component), which utilised the advantages of both the expert rules extracted automatically from coding guides and the labeled data:

Extended rule-based system: Since missing transliterations and synonyms can be captured through the false negative predictions of the system, we decided to build statistical models to learn to predict the false negatives of our ICD-9-CM coder. This way we expected to have the most characteristic phrases for each label among the top ranked features for a classifier model which predicted the false negatives of that label.

Extended classification system: We can import the rule-based system into the classification model by incorporating its predictions into the feature space of the latter. We added all the codes predicted by the rule-based system to the Vector Space Model representation. Thus the statistical system can exploit the knowledge of both the coding guides and the regularities of the labeled data.

Hybrid model: One of the most straightforward approaches for the combination in this multi-labeling environment is to take the union of the labels predicted by the rule-based expert system and the Machine Learning model. In this setting we made predictions using the expert system and the classifier quite independently.

The CMC challenge itself was dominated by entirely or partly rule-based systems that solved the coding task using a set of hand crafted expert rules. The feasibility of the construction of such systems for thousands of ICD codes is indeed questionable. Our results demonstrate that hand-crafted systems – which proved to be successful in ICD-9-CM coding – can be reproduced by replacing several laborious steps in their construction with statistical learning models [14].

Classifying patient records whether they have a certain disease is a similar task to ICD coding. The *Obesity Challenge* in 2008, organised by the Informatics for Integrating Biology and the Bedside

(I2B2)[1], asked participants to construct systems that could correctly replicate the textual and intuitive judgments of the medical experts on obesity and its co-morbidities based on narrative patient records [15].

Our textual classification approach focused on the rapid development of an extended dictionary-lookup-based system, which also took into account the document structure and the context of disease terms for classification. To realise this, we used statistical methods to pre-select the most common (and most confident) terms and abbreviations then evaluated outlier documents to discover infrequent terms and spelling variants (data-driven approach) and extended the dictionaries by lists from the MedlinePlus encyclopedia[2] [16].

Exploiting non-textual relations among documents

In an applied Information Extraction task the documents to be processed are usually not independent of each other. The relations among documents can be exploited in the IE task itself. We introduced two tasks where graphs are constructed and employed based on these relations. In the biological *Gene Name Disambiguation* task we utilised the co-authorship graph, while in the *Opinion Mining* task the response graph was built.

Biological articles provide a huge amount of information about genes, proteins and their behaviour under different conditions. The task of biological entity recognition is to identify and classify gene, protein, chemical names in biological articles [17]. Taken one step further, the goal of Gene Name Normalisation (GN) [18] is to assign a unique identifier to each gene name found in a text. One gene name can refer to different entities (for example, *IL-21* can refer to the genes with EntrezGeneID 27189, 50616 or 59067). Gene Symbol Disambiguation (GSD) [19] is a subtask of GN whose goal is to select the correct sense – the gene ID from a well-defined inventory – of a gene name according to its context.

Our main idea here was that an author habitually uses gene names consistently; that is, they employ a gene name to refer exclusively to one gene in their publications [20]. Generalising this hypothesis we may assume that the same holds true for the co-authors of the biologist in question. But what is the situation for the co-authors of the co-authors? To answer this question - and utilise the information obtained from co-authorship in the GSD problem - we decided to use the so-called co-author graph [21]. Several ways of obtaining information from co-authorship along with textual information were utilised and introduced for solving the GSD task.

We exploited graphical information for an *Opinion Mining* (OM) task [22] as well. OM seeks to extract opinions and polarity about a certain topic from unstructured texts. This task has been attracting increasing academic interest in NLP for over a decade. This phenomenon is due to the fact that people nowadays are more likely to share their emotions and opinions toward various topics. From these rich sources of opinions valuable information can be extracted, which can help, for instance, political parties to design their campaign programme or companies to get feedback about their products based on opinions expressed on the internet.

Here our hypothesis was that people representing different views in the debate would comment more frequently on each other's posts compared to others. Thus, we composed a weighted, directed graph, in which each node is mapped to a person and the weight of an edge(A, B) corresponds to the number of person B's replies towards person A. We obtained this information from the HTML structure of the pages, but it is worth mentioning that not everyone indicated whether they were replying to another post and some people did not use this feature correctly. In our final system we combined the results of two Machine Learning methods in our system. One of them was based on the traditional Vector Space Model while the other one was trained on data derived from the so-called interaction or *response graph*. Empirical results are presented on classifying the member of a Hungarian forum based on their opinions about a topic of the forum.

[1] www.i2b2.org/NLP/

[2] http://www.nlm.nih.gov/medlineplus/encyclopedia.html

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.

- Part I – Supervised learning for Information Extraction tasks
  - Chapter 3: **Supervised models for Information Extraction**

    For NER in Hungarian the author participated in the creation of the first Hungarian NE reference corpus [7], which allowed researchers to investigate statistical approaches. Together with his colleagues, the author constructed a Machine Learning-based NER system [2] and a metonymy resolution system (GYDER) [4]. Our classification systems achieved results on international reference datasets which were competitive with other state-of-the-art NER taggers. The construction of the SzegedNE corpus was an inseparable contribution of the authors of [7].

    The major contribution of the author here is an experimental comparison of the token-based versus sequential approaches, and several classification algorithms. He gave several suggestions about which algorithm should be used in a given learning environment as well. The author argues that IE tasks share the same common characteristics and NER (along with metonymy resolution) can be regarded as a prototype task, so the conclusions drawn from it can be generalised to other tasks. When turning to the issue of making a comparison several points arise:

    * The huge amount of discrete features in Information Extraction tasks imply the use of decision trees or generative models.
    * The decision tree has the most favourable time complexity.
    * Only the output of the decision tree is directly interpretable by users.
    * The generative probabilistic models – like Logistic Regression and CRF – output a probability distribution on the class labels which could be a confidence measure for an application.

    Overall, we recommend using decision trees in the development phase (experiments) of an application because of its training time, ease of interpretability and use of generative models (Logistic Regression in classification and CRF in sequence labeling tasks) in the final versions.

  - Chapter 4: **Combinations of learning models**

    The author presented experiments on combination schemes of different Machine Learning algorithms. The building and testing of less frequently applied algorithms is always worth doing, since they can have a positive effect when combined with popular models. The author empirically verified this statement in the Hungarian and English NER tasks, applying Stacking [2][3], Boosting [3] and co-training [23]. He presented a novel combination approach, called *Feature set split and recombination* [3], which exploits the rich feature set.

    These combination schemes play a key role in the construction of the complex NER system by the author with his colleagues. This NER system is competitive with the published state-of-the-art systems (an F-measure of 89.2% and 94.77% for English and for Hungarian, respectively). It has a different theoretical background compared to the widely used sequential ones, which makes it an excellent candidate for a combination scheme with external NER systems. The main result here can summarised as follows. Simple combination schemes are worth employing because they usually bring about a significant accuracy improvement. Moreover, the scores are usually better than those achieved by employing more sophisticated but time-consuming stand-alone learning algorithms.

– Chapter 5: **On the adaptability of IE systems**

Together with his colleagues, the author participated in the 2006 I2B2 shared task challenge on medical record de-identification [5] with a (domain-)adapted version of the pre-existing NER system. The major steps of the adaptation, and results achieved (as a whole) are the joint contribution of the co-authors. As our results clearly show, the system we obtained via the domain adaptation of our newswire NER model is competitive with other approaches and achieved the best scores in phrase-level evaluation among the systems submitted to the challenge, without any statistically significant difference in performance from the other top-performing system.

In [10] the author's contribution is the general idea of using NER in sentence alignment systems, while the other general concepts of sentence segmentation and alignment were actually carried out by the co-authors. In [11] the author performed several inter-domain statistical investigations on the BioScope subcorpora.

The key results here might seem a bit strange at a first glance and can be summarised as follows. In the middle application layer of IE systems like NER – where the deep language-specific facts (like morphological and POS codes) are encoded into features – the statistical systems work language independently, while changing the domain in a certain language requires much more effort to achieve the satisfactory level of performance.

- Part II – **Exploitation of external knowledge in Information Extraction tasks**

  – Chapter 6: **Using the WWW as the unlabeled corpus**

  The author with his colleagues developed WWW-based NER post-processing heuristics and experimentally investigated them on general reference NE corpora [24]. They constructed several corpora for the English and Hungarian NE lemmatisation and separation tasks. The NE lemmatisation task is important for textual data indexing systems, for instance, and is of great importance for agglutinative languages like Finno-Ugric and Slavic languages. Based on these constructed corpora machine learnt decision rules were introduced [25]. These solutions are based on the assumption that, even though the World Wide Web contains a good deal of useless and incorrect information, for our simple features the frequency of correct language usage dominates misspellings and other sorts of noise.

  The author's own contributions in the Web-based solutions are the most frequent role and phrase extension approaches in [24]. The feature engineering tasks and most of the Machine Learning experiments of [25] were carried out by the author.

  – Chapter 7: **Integrating expert systems into Machine Learning models**

  The author with his co-authors developed solutions for clinical IE tasks [14; 16] that integrate Machine Learning approaches and external knowledge sources. They exploited the advantages of expert systems and statistical models. Expert systems are able to handle rare labels effectively. Statistical systems on the other hand require labeled samples to incorporate medical terms into their learnt hypothesis and are thus prone to corpus eccentricities and usually discard infrequent transliterations, rarely used medical terms or other linguistic structures.

  Overall, we think that our results demonstrate the real-life feasibility of our proposed approach and that even systems with a shallow linguistic analysis can achieve remarkable accuracy scores for information extraction from clinical records.

  Each statistical system along with the above-mentioned integration methods developed for the two tasks [14; 16] were the author's own contributions.

  – Chapter 8: **Exploiting non-textual relations among documents**

  The author examined the utility of graphs based on relations among documents in Information Extraction systems. He experimentally demonstrated the utility of co-authorship analysis for the GSD task [20] and achieved an outstanding accuracy (97.22% precision

at 100% recall). His hypothesis was that a biologist refers to exactly one gene by a fixed gene alias, and in experiments we found evidence for this.

The author with his colleagues developed an Opinion Mining system which uses the information gathered from texts along with the response graph [26]. For this task the first corpus dedicated to Opinion Mining in Hungarian was constructed and results close to the inter-annotator agreement rate were achieved.

All the contributions described in [20] are independent results of the author. In [26], the author's contribution is the idea and general concept of using the response graph for Opinion Mining.

			Chapter					
			3	4	5	6	7	8
ACTA	2006	[2]	•	•				
LREC	2006	[7]	•					
SEMEVAL	2007	[4]	•					
DS	2006	[3]		•				
JAMIA	2007	[5]			•			
BIONLP	2008	[11]			•			
ICDM	2007	[24]				•		
TSD	2008	[25]				•		
BMC	2007	[14]					•	
JAMIA	2009	[16]					•	
BMC	2008	[20]						•
DMIIP	2008	[26]						•

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

Summary by papers

Table 1 summarises the relationship among the thesis chapters and the more important[3] referred publications of the author. Here we list the most important results in each paper that are regarded as the author's own contributions. We should 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 exceptions are [20], which describes only the author's own results and [7] as all the results described in this paper are counted as shared contributions of the authors. For [4], the author made only marginal contributions.

- ACTA 2006 [2]
  - Comparison of supervised learners.
  - Stacking approach.
- SEMEVAL 2007 [4]
  - Comparative analysis of C4.5 and Logistic Regression.

[3]For a full list of publications, please visit http://www.inf.u-szeged.hu/~rfarkas/publications.html.

- DS 2006 [3]
  - The architecture of the complex NER model.
  - The feature set split and recombination method.
  - Boosting experiments.
  - Post-processing rules.
- JAMIA 2007 [5]
  - Trigger-based bagging method.
  - The standardisation phase.
- BIONLP 2008 [11]
  - Statistical investigations for differences between the medical and biological domains.
- ICDM 2007 [24]
  - Using web frequencies for phrase boundary extension.
  - The most frequent rule heuristic.
- TSD 2008 [25]
  - Feature set construction and transformations for NE lemmatisation and separation.
  - All of the Machine Learning experiments.
- BMC 2007 [14]
  - Combination strategies for expert rules and Machine Learning methods.
  - Negation and speculation-based language pre-processing.
  - Multi-label classification approaches.
  - All of the data-driven experiments.
- JAMIA 2009 [16]
  - Statistical methods for term identification.
  - Statistical methods for context detection.
- DMIIP 2008 [26]
  - The general idea of using response graphs for Opinion Mining.

References

[1] Wolpert DH, Waters R: **The relationship between PAC, the statistical physics framework, the Bayesian framework, and the VC framework**. Tech. rep., The Santa Fe Institute 1994.

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

[3] Szarvas Gy, Farkas R, Kocsor A: **A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms**. *DS2006, LNAI* 2006, **4265**:267–278.

[4] 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].

[5] Szarvas Gy, Farkas R, Busa-Fekete R: **State-of-the-art anonymisation of medical records using an iterative machine learning framework**. *Journal of the American Medical Informatics Association* 2007, **14**(5):574–580, [http://www.jamia.org/cgi/content/abstract/M2441v1].

[6] Markert K, Nissim M, Lw BPE: **Metonymy resolution as a classification task.** In *Proceedings of EMNLP* 2002:204–213.

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

[8] Sang TK, F E, 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.

[9] 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].

[10] Tóth K, Farkas R, Kocsor A: **Sentence Alignment of Hungarian-English Parallel Corpora Using a Hybrid Algorithm.** *Acta Cybernetica* 2008, **18**(3):463–478.

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

[12] Halácsy P, Trón V: **Benefits of Resource-Based Stemming in Hungarian Information Retrieval**. In *CLEF* 2006:99–106.

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

[14] 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].

[15] Uzuner O: **Recognizing Obesity and Co-morbidities in Sparse Data.** *Journal of American Medical Informatics Association* 2009.

[16] Farkas R, Szarvas Gy, Hegedűs I, Almási A, Vincze V, Ormándi R, Busa-Fekete R: **Semi-automated construction of decision rules to predict morbidities from clinical texts.** *Journal of the American Medical Informatics Association* 2009, **accepted for publication**.

[17] Yeh AS, Hirschman L, Morgan AA: **Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup**. *CoRR* 2003, **cs.CL/0308032**, [http://dblp.uni-trier.de/db/journals/corr/corr0308.html#cs-CL-0308032].

[18] Hirschman L, Colosimo M, Morgan A, Yeh A: **Overview of BioCreAtIvE task 1B: normalized gene lists**. *BMC Bioinformatics* 2005, **6**(Suppl 1):S11.

[19] Xu H, Markatou M, Dimova R, Liu H, Friedman C: **Machine learning and word sense disambiguation in the biomedical domain: design and evaluation issues**. *BMC Bioinformatics* 2006, **7**:334, [http://www.biomedcentral.com/1471-2105/7/334].

[20] Farkas R: **The strength of co-authorship in gene name disambiguation**. *BMC Bioinformatics* 2008, **9**, [http://dx.doi.org/10.1186/1471-2105-9-69].

[21] Barabasi AL, Jeong H, Neda Z, Ravasz E, Schubert A, Vicsek T: **Evolution of the social network of scientific collaborations**. *Physica A: Statistical Mechanics and its Applications* 2002, **311**(3-4):590–614.

[22] Ghose A, Ipeirotis PG, Sundararajan A: **Opinion Mining using Econometrics: A Case Study on Reputation Systems**. In *ACL* 2007.

[23] Farkas R, Szarvas Gy, Csirik J: **Special Semi-Supervised Techniques for Natural Language Processing Tasks.** In *Proceedings of the 6th International Conference on Computational Intelligence, Man-Machine Systems and Cybernetics* 2007:360–365.

[24] Farkas R, Szarvas Gy, Ormándi R: **Improving a State-of-the-Art Named Entity Recognition System Using the World Wide Web**. *ICDM2007, LNCS* 2007, **4597**:163–172.

[25] Farkas R, Vincze V, Nagy I, Ormándi R, Szarvas Gy, Almási A: **Web based lemmatisation of Named Entities.** In *Proceedings of the 11th International Conference on Text, Speech and Dialogue* 2008:53–60.

[26] Berend G, Farkas R: **Opinion Mining in Hungarian based on textual and graphical clues.** In *Proceedings of the 4th International Symposium on Data Mining and Intelligent Informaion Processing* 2008.