Multi-domain Multi-lingual Named Entity Recognition: Revisiting & Grounding the resources issue

Voula Giouli, Alexis Konstandinidis, Elina Desypri, Harris Papageorgiou

ILSP/IRIS
Artemidos 6 & Epidavrou, Maroussi, Greece
{voula, alex, elina, xaris@ilsp.gr}

Abstract

The paper reports on the development methodology of a system aimed at multi-domain multi-lingual recognition and classification of names in texts, the focus being on the **linguistic resources** used for training and testing purposes. The corpus presented here has been collected and annotated in the framework of different projects the critical issue being the development of a final resource that is **homogenous**, **re-usable** and **adaptable** to different domains and languages with a view to robust multi-domain and multi-lingual NERC.

1. Introduction

Language Resources (LRs) production has always been considered a critical issue and a requisite for the development of successful Human Language Tecnology applications, especially when state-of-the-art techniques are employed, that require large corpora coupled with metadata for training and testing purposes. Moreover, tools that have been developed for one language or domain need extensive training prior to being ported into other languages and/or new domains. This paper reports on work in progress intended for the construction of multi-lingual and multi-domain LRs that were initially collected and annotated in different settings, and on our efforts for their successful harmonisation, so that they are rendered homogenous and, thus, re-usable. The corpora were designed to assist in the training and testing process of a system aimed at spotting and classification of Named Entities (NEs) developed at the Institute for Language and Speech Processing (ILSP) in the framework of the MUSE project. The system is language and domain independent, yet it was initially trained and evaluated on Greek news politics data. Within the framework of Reveal-This project, however, adaptation of the system and accompanying LRs to other languages (English and French) and domains (travel, politics, health) is now being attained.

2. NERC across languages and domains: the problem of customization

Named Entity Recognition (NERC) consists in the identification and classification of phrases denoting certain entities of given categories (locations, persons, organizations, etc.), that are of importance to a range of applications. Having a long-established experience in the development of NLP tools with a view to Information Retrieval and Extraction, ILSP had initially implemented a rule-based approach to NERC for the Greek language (Demiros et al. 2000) that incorporated deep linguistic knowledge. System development was performed on the basis of a financial corpus of Greek texts collected from various sources over the internet and tagged in accordance with the MUC (Message Understanding Conference) specifications. Evaluation of the system on the financial documents proved it to be successful (F=83%) outperforming similar tools for the English language.

However, when having to cope with new domains and/or languages as in the framework of MUSE and Reveal-This projects, the system yielded poor results. MUSE was a nationally-funded project aimed at the development of a robust system for the analysis, annotation, archiving, indexing and retrieval of large amounts of audiovisual content related to business news, the final purpose being the personalized delivery of content as well as associated metadata over multiple devices. And, whereas MUSE was focused on Greek data only posing the problem of system portability to a new domain, in the framework of Reveal-This the issue of multi-lingual multi-domain NERC needed to be coped with. Moreover, the new domains, namely, travel, politics, health, and the applications at hand required a much more sophisticated NE classification schema than the one already adopted in the framework of MUSE and our previous rule-based approach. Therefore, not only our our methodology with respect to NERC had to be revisited due to the costs entailed by the adaptation process, but also our specifications with respect to NE categorization had to be updated, the critical issue being the development of resources that are homogenous, re-usable and adaptable to different domains and languages with a view to robust multi-domain and multi-lingual NERC.

3. Methodology revisited

Grammar development is a time consuming and costly effort. Moreover, when system portability to new languages and domains is required, more robust methods should be investigated.

3.1. State-of-the-Art

Our approach to NERC capitalizes state-of-the-art techniques and builds on the methodologies presented and evaluated in two consecutive conferences on Computational Natural Language Learning (CoNLL), namely systems that participated in the CoNLL-2002 and CoNLL-2003 shared task. Both CoNLL-2002 and ConLL-2003 involve language-independent NERC, yet, practically each focuses on two languages only: CoNLL-2002 focus was on Spanish and Dutch, whereas participants of ConLL-2003 have been offered training and test data for English and German. For each of the languages there has been a training file, a development file and a test file. ConLL-2003 has also catered for a large file with unannotated data to be incorporated in the learning process. The learning methods were trained with the training data. The development data were used for tuning the parameters of the learning methods. The following types of names were dealt with: persons (PER), organisations (ORG), locations (LOC) and miscellaneous names (MISC) that do not fall into the other three categories. Performance of the participating systems was measured with the recall (R), precision (P), and F-measure (F = 2PR / (P+R)) scores. The most frequently applied technique with top results in one or both languages in the CoNLL-2003 shared task was the Maximum Entropy Model. (Bender et al., 2003); (Chieu and Ng, 2003); (Curran and Clark, 2003) used Maximum Entropy Model in isolation, whereas, others (Florian et al., 2003; Klein et al., 2003) used ME methods in combination with other techniques.

3.2. NERC revisited: MENER

MENER, our NE recognition system is a highly modified version of the best-scoring system in CoNLL-2003 shared task developed by Hai Leong Chieu and Hwee Tou Ng. It is a single-level maximum entropy approach to NERC that makes use of a broad range of features extending from conjunctional ones, that lend the system limited pattern recognition abilities, to individual ones, that indicate statistically important evidence extracted automatically from the training data. The main idea was to maximize the probability $p(N|S,Doc)$, where N is the sequence of NE tags assigned to each word in sentence S, and Doc is the relevant information that can be extracted from the whole document containing S. To this end, sentence-based local evidence about words is combined with global evidence, the latter being a collection of features drawn from other occurrences of these words within the same document (global features).

Other machine learning—based NERC systems usually try to maximize the probability $p(N|S)$ only, or, often, make use of global data by incorporating a second-level classifier that tries to improve the output of its sentence-based predecessor.

The publicly available (LGPL licensed) Java based OpenNLP maximum entropy package (maxent) has been employed in our implementation. Using a sliding window of four tokens ($[\text{token}]_i$ or focus token, the preceding ($[\text{token}]_{i-1,}$ and the next two tokens – should they exist), the NE-Tagger scans the entire input XML file. It collects all contextual features applicable to the focus token by consulting search and cache units appropriately along with the system's resources and submits the resulting set of features to the ME model for evaluation. The response of the ME model is a probability distribution over the classes it has been trained on. Accordingly, NE-Tagger selects the class with the maximum probability to be assigned to the focus word and restarts the processing cycle by sliding the window to the next token.

Additionally, in an effort to circumvent the lack of memory that is inherent in every ME model by injecting its previous decisions, as an additional feature, into contexts that will be evaluated in the future, the system cache unit accumulates already recognized NE tokens along with the tags assigned. The cache unit is empty only when processing reaches the end of each tagged document. The search unit is actually the provider of global features. Using document-wide searches, it examines, for each token, the context of its other occurrences looking for trigger words and other clues (e.g. capitalisation information). Extracted features from each occurrence are united to form a set, which, as it grows, keeps only the features with high ranking, in order to eliminate weak evidence. All the features contained in this set have an extra prefix that emphasizes their origin, in order to render them distinguishable from their local context counterparts. After all occurrences are exhausted, the unit returns the resulting feature set.

3.3. MENER linguistic resources

The system is also coupled with linguistic resources that have been automatically extracted from the training data during the pre-training stage. For each name class the following set of lists has been compiled:

- **Unigrams**: single words that precede the name class,
- **Bigrams**: bigrams of words that precede the name class. In order to keep the strongest evidence only, this list includes bigrams with higher probability to appear before the name class than the contained unigram itself (for example, "city of" vs. "of", the first one appears more often before locations than the other),
- **Post unigrams**: words that succeed the name class
- **Suffixes**: three letter suffixes of words pertaining to the name class,
- **Prefixes**: three letter prefixes of words pertaining to the name class,
- **Terminal tokens**: tokens that terminate the name class, for example the organization class often terminates with tokens such as "Inc." and "Corp.",
- **Functional words**: lower case words or punctuation symbols that occur within the name class, for example "van der", "&", "of", etc.

Apart from name classes, the system also consults a **Frequent Word List (FWL)** that consists of words occurring in the training data over a given threshold. This is set to 4 for the EN data and 3 for the EL model, and is used to determine the rareness of a word, a fact that is reflected as an extra feature. Additionally, MENER may, optionally, make use of an external knowledge base in the form of lists with line-delimited records of **known names** (per name class), compiled from a variety of sources (Internet, CoNLL 2003 shared task data, annotated data, etc.).

Apart from FWL and name lists, all other lists are sorted according to the ascending order of the *correlation coefficient* (Ng et al, 1997) C of an item w in relation to a name class NC, which is defined as:

$$C = \frac{(N_{r+}N_{n-} - N_{r-}N_{n+})\sqrt{N}}{\sqrt{(N_{r+} + N_{r-})(N_{n+} + N_{n-})(N_{r+} + N_{n+})(N_{r-} + N_{n-})}}$$

where N is the total number of sentences in training data, N_{n-}, N_{n+}) is the number of non-relevant sentences (in which the item w does not occur) that contain no (at least one) token of NC class, N_{r+} and N_{r-} refer to the number of relevant sentences which do include item w either under right conditions that meet its meaning or not, respectively. For example, in case of a unigram or bigram item w, N_{r+} refer to number of sentences in which w actually precedes an instance of NC class. Correlation coefficient is a variant of x^2 metric and can be characterized as a "one-sided" x^2 metric. It selects exactly those items that are highly indicative of membership in a category, whereas x^2 will also pick items that are indicative of non-membership in the category.

4. NERC corpus: the need for harmonizng resources

Training of MENER has been performed on the basis of a corpus that originated from different projects: (a) MUSE that aims at Greek NERC in the domain of news politics and (b) REVEAL-This which focuses on three languages: Greek, English, and French and on the domains of news, politics, travel and health. The inventory of ILSP also comprises a Greek NE MUC-conformant corpus that pertains to the financial domain (Demiros et al., 2000). All corpora have been annotated at various levels of linguistic analysis according to the needs and requirements imposed by the applications at hand.

The MUSE corpus originated from the the Greek Business Channel (GBC) and amounts to 300k. GBC data consists of short daily news bulletin files, covering economical, domestic and world news, which briefly mention 6-8 events per file. Despite their size, they are full of named entities and hence are valuable for the NERC task. The corpus was initially annotated with accordance to the MUC annotation schema catering for the following types of NEs: *PERSON*, *ORGANISATION* and *LOCATION.*

Reveal-This, on the other hand, covers three languages (EL, EN, FR) and the domains of news, politics, travel and health. The project is still on-going, and for the time being, only Greek and English have been accomodated for in the domains of news, politics and travel. To this end, a comparable corpus in the afore-mentioned languages has been collected that comprises both written textual data and transcriptions generated from audio-visual content. More specifically, the Greek news corpus is an agglomeration of mainly news documents, exclusively collected for the Reveal-This project from the following sources: various Internet news sites (9k tokens), the Greek Business Channel (GBC) (107k) and data that originate from the European Parliament (EP) web site (66k). EP files are normalized transcripts of European Parliament sessions. Its English counterpart comprises textual web data (9k) and EP documents (54k). It has also been coupled with the English data provided at the CoNLL-2003 shared task. The material is part of the Reuters corpus. It consists of three parts: the core data (204k tokens), a development set (51k) and a test set (46k), all provided in a line-delimited textual format that we converted into an equivalent XML representation. The travel data amount to 72k tokens for the Greek data and 47k for its English counterpart. This corpus has been annotated according to the Automatic Content Extraction (ACE) specifications.

5. The NE Annotation Schema

In our initial efforts with respect to NERC, guidelines pertaining to MUC were adopted due to the fact that they seemed more appropriate for the domain chosen and the application at hand, that is, information retrieval and extraction from Greek financial texts. According to these guidelines the following types of NEs were spotted and classified: person (*PER)*, organization (*ORG)* and location (*LOC).* The schema also caters for the identification of numerical values: *MONEY, PERCENT,* and certain time expressions: *DATE* and *TIME.* This schema, however, was proved to be inadequate for the efficient handling of new domains such as politics, travel and health, as distinctions were not fine-grained and ambiguity was carried on to be resolved at later stages. Moreover, the application at hand that went beyond the mere filling of templates for information retrieval and extraction purposes, asked for more complicated semantic information. The purpose being to semantically enrich the contents of multilingual multimedia documents with topic, entity and fact information relevant to user profiles, a more sophisticated annotation schema should be chosen.

To fill the gap, we have opted for a classification schema that was compatible with current trends in NERC, namely the ACE (Automatic Content Extraction) schema catering for the recognition and classification of the following types of NEs: person (*PER)*, organization (*ORG)*, location (*LOC)* and geopolitical entity (*GPE*). Moreover, NE's of the type *LOC* were also assigned a subtype value, namely: location (LOC) geographical region (GEO) and facility (FAC). Though compatible in form with ACE, in that it retains most of the types and subtypes provided for by ACE, our classification schema differs in that disambiguation between *LOC* and *GPE* uses of names is being attempted. To sustain, on the other hand, compatibility with the ACE schema, we have catered for an extended annotation schema with subtypes for further classifying the spotted NE's. Due to the type and content of the project data, however, only subtypes of LOC entities have been used. Moreover, due to limitations imposed by the data for the French NERC that employ a rather flat, MUC-conformant schema catering for only PER, ORG and LOC entities, a condensed classification schema has also been retained. To this end, mappings of the extended classification schema have been performed: GPE has been mapped on LOC, and LOC sub-classification has been dropped. A short description of our NE extended annotation schema and relevant guidelines is provided hereby:

PERSON: Names of individuals, family names and widely used aliases or nicknames of people are marked as NE's of the type PERSON. Similarly, proper names that refer to saints or dead people are also marked as PERSON NE's unless they are used to name other entities (i.e., ships, churches, locations, prizes or awards, etc.). Within the current schema, occupations, titles, honorific expressions that usually precede a name are not considered as part of the markable NE, e.g.:

President [person Borrell Fontelles /person]

The annotation schema has also provided for the following subtypes of the type PER though they have not been applied on the data:

- PER.human: Names of people, either dead or alive are further classified as human: *Mr Ortuondo Larrea*
- PER.animal: Names of animals fall into this subtype: *Morris the cat.*
- PER.fictional: Names of fictional characters are tagged: *Spiderman is children's hero.*
- PER.other: All other animate entities that do not fall into the above subtypes are to be tagged as PER.other.

ORGANISATION: Companies, enterprises, organizations or groups of people with an organizational status fall within this category and are marked as NE's of the type ORGANISATION, e.g., :

the [org Iraqi government /org]

On behalf of the Group of the [org European People's Party /org] and [org European Democrats /org]

The annotation schema has also provided for the following subtypes of the type ORG though they have not been applied to the data:

- ORG.commercial: A commercial organization is focused primarily upon providing ideas, products, or services for profit, such as industries, industrial sectors, etc.
- ORG.educational: Institutions focused primarily upon the furthering or promulgation of learning fall into this sub-class.
- ORG.other: All other organizations that do not fall into the above subclasses.

LOCATION: Proper names that designate landmarks are marked as being of the type LOCATION, e.g.:

[loc Poland /loc] was also the birthplace of the revolt against totalitarianism.

The following subclasses have been used at the annotation:

- LOC.geo:Geographical entities, that have been created naturally upon or above the surface of the earth, such as mountains, masses of water, etc.
- LOC.loc: Geographical regions that do not pertain to the above class. Contextual information is used to distinguish a LOC.loc from a GPE entity.
- LOC.fac: Large functional man-made constructions are facilities, that is artifacts that fall under the domain of architecture and civil engineering. Contextual information is used to distinguish a LOC.fac from an ORG entity.
- LOC.other: Other entities that are used to designate a space fall into this class, such as stars, planets, etc.

GEOPOLITICAL ENTITY: Geopolitical Entities (GPE) are geographical regions also defined by political and/or social groups. According to ACE specifications "A GPE entity subsumes and does not distinguish between a nation, its region, its government, or its people". In our schema, however, context has guided disambiguation between LOC and GPE uses of names:

I especially want to welcome the arrival of [gpe Cyprus /gpe]

She visited [loc Cyprus /loc]

GPE entities are further classified with the following subtypes of the type GPE though they have not been applied to the data:

- GPE.continent
- GPE.nation
- GPE.province
- GPE.city
- GPE.other

It should be noted that markable entities appear in the text with their full-name, an abbreviated/reduced form of this name, or a word/phrase - usually a metonymy - consistently used to describe it, and all these alternative mentions are tagged. However, simple pronominal or nominal references to NEs are not marked.

[org National Bank of Greece /org] – [org Εθνική / National /org]

[org Ηρακλής /org] (=soccer team) - ο [org Γηραιός / org]

[org Athens Stock Exchange /org] – [org ASE /org]

President [person Borrell Fontelles /person] – the president said (it is not marked)

Words that usually precede a name (articles, modifiers, etc) are not to be included within the markable NE:

the [org Iraqi government /org]

It should be noted that NEs that are connected through part-whole and possessor-possessed relations are not marked as a single entity:

Το [org Τμήμα Ανάλυσης και Μελετών /org] της [org Εγνατίας ΑΧΕ /org] / The [org Research Department /org] of [org Egnatia Securities /org]

Initial guidelines were provided by the linguists that perform the annotation task. After a brief testing period, samples by all members of the team of annotators were collected and inter-annotator agreement was examined. The guidelines were further augmented with cases that linguists consider exceptional, or where systematic inter-annotator disagreement had been observed.

Annotation follows the so-called IOB-1 format, in which an I-tag (e.g. I-person) is used for all words in an entity, including the first word, unless the first word separates contiguous entities of the same type, in which case a B-tag (e.g. B-loc) is used. All words outside a named entity are marked with the O tag.

6. Marker: A GUI for multi-level corpus annotation

Annotations have been carried out by means of a GUI developed at ILSP called *Marker*. Marker is an environment that allows annotators to have simultaneous views of all levels of previous annotations, while working at a particular task. Furthermore, it is equipped with comparison facilities that allow for inspection of interannotator agreement or tool performance, expressed in precision and recall measures. The environment currently supports annotation at the morphosyntactic level, chunk and recursive phrases level, NE, term and coreference annotation, and annotation of grammatical relations. The tool runs on any PC or workstation equipped with a recent version of Sun's Java 2 Runtime Environment and is available free of charge for research purposes.

The environment also provides a text box for the creation and editing of comments, which are stored inside the metadata files, as child elements of the relative <sent> element. Other options include choice of annotation level, expanding and collapsing trees (when editing large sentences), etc. All preferences are stored in user-profile files and can be retrieved each time the tool is run.

Moreover, session metadata are stored separately from annotation data. These metadata elements are inspired by the Dublin Core Metadata Initiative (DCMI) standard, including among others, *Annotator* (an entity responsible for providing the annotation content), *Subject* (what the annotation is about), *Resources* (the resources that have been used in the annotation session), *Language* and *Date* (a date associated with the current session). Subsequent modifications/reviews by the same or other annotators are also kept in the session metadata files. Classes of XML annotations that share a common vocabulary and structure (morphology, syntax, etc.) are described in DTD's. The Marker looks for the relevant DTD when initiating an annotation session and configures the GUI appropriately by providing the needed functionality to the annotator. This dynamic process of building and customising a GUI on the fly (based on external DTD files) is currently restricted to simple elementary structures which however fulfill most of our current annotation needs. Additionally, a validation step is being performed ensuring that a particular instance is compliant with the prespecified constraints in the DTD's. Annotation files produced with Marker are closely coupled with their corresponding source files, in terms of referential links, but cannot be feed directly into MENER. So, a separate tool carries out the reference resolution and embeds the links back into the source files using tags that comply with the IOB-1 format.

7. System Evaluation

Evaluation has been performed on the basis of the condensed tagset. Our system outperformed the system developed by Chieu and Ng for ConLL2003 in the Greek Politics and News Politics data rendering an F-measure of 94.87 (see Table 2). Yet it performed rather poorly in the English data that pertain to the same domain and text type (see Table 1), ORG class being the most problematic. This is actually attributed to the fact that the English model was extensively trained on the Reuters data, whereas evaluation was performed on the REVEAL-This data only. And, whereas data collected from web resources are similar in structure to the Reuters data set, texts originating from the Europarliament exhibit peculiarities (i.e., text and period structure, headlines and capitalisation conventions, etc.) that were not taken into account at the training phase. In the following examples, the system has erroneously recognised the words *Members* and *Rules* as being a NE of the type ORG:

... being very similar to the vote of the <org Members /org> of the <org European Parliament /org>

...pursuant to <org Rules /org> 130 and 131 of the <org Rules /org> of Procedure

As far as the travel data are concerned, our system performs well in PER and LOC classes, but renders poor results in class ORG (see Table 3 and Table 4). This is attributed to the fact that the texts at hand allow for a large degree of ambiguity between ORG-PER and ORG-LOC that cannot easily be resolved solely on the basis of statistical information, and further contextual knowledge should be taken into account:

Vasilis and Eliza Goulandris Foundation organizes every summer at the <loc Museum of Modern Art /loc> exhibitions

The <org Museum of Modern Art /org> in Andros organized the exhibition.

Quantitative results are given in the tables below:

Politics_ EN	**NE**	**Precision** (%)	**Recall** (%)	**F-measure**
Training : ConLL data	loc	69.15	84.75	76.16
Testing: Web + EP	org	47.14	87.54	61.28
Iterations/ Cut-off: 400/1	per	81.37	73.38	77.17
	Total	**59.17**	**82.70**	**68.98**

Table 1: Results for the EN politics and news politics data

News_EL Evaluation	**NE**	**Precision** (%)	**Recall** (%)	**F-measure**
Training : EP + GBC + Web	loc	95.88	93.40	94.62
Testing : EP + GBC + Web	org	90.65	94.61	92.59
Iterations/Cut-off: 400/1	per	99.02	98.54	98.78
	Total	**94.82**	**94.92**	**94.87**

Table 2: Results for the EL politics and news politics data

Travel_EN Evaluation	**NE**	**Precision** (%)	**Recall** (%)	**F-measure**
Training : Reveal	Loc	69.68	78.97	74.04
Testing : Reveal	Org	20.00	14.29	16.67
Iterations : 350	Per	63.33	65.52	64.41
Cut-off : 1	Total	**67.97**	**75.32**	**71.46**

Table 3: Results for the EN Travel data

Travel_EL Evaluation	**NE**	**Precision** (%)	**Recall** (%)	**F-measure**
Training : Reveal	Loc	71.97	87.37	78.93
Testing : Reveal	Org	35.00	25.00	29.17
Iterations : 400	Per	66.67	43.14	52.38
Cut-off : 1	Total	**70.23**	**78.80**	**74.27**

Table 4: Results for the Greek Travel data

8. Conclusions

We have presented a resource that was constructed out of different subcorpora and that was used to guide training and testing of multi-lingual multi-domain NERC. To this end, harmonisation of the respective corpora was carried out by specifying a new annotation schema that is compatible to new trends. Annotations were applied to all data, with mappings that allow previous annotations with a more flat representation to be retained.

9. Acknowledgements

We would like to thank three anonymous reviewers for useful suggestions and comments. Work described in this paper was fully supported by the research project “Multimedia Content Management Systems” (MUSE), funded in the framework of Axis 3, Measure 3.3 of the Concerted Programme for Electronic Business of the General Secreteriat for Research and Technology of the Greek Ministry of Development as also by the research project "Retrieval of Video and Language for The Home user in an Information Society" (REVEAL THIS), FP6-IST-511689, funded in the framework of the specific research and technological development programme "Integrating and strengthening the ERA".

10. References

Annotation Guidelines for Entity Detection and Tracking (EDT) Version 4.2.6 20040401.

Bender, O., Och F.J., and Ney, H. (2003). Maximum Entropy Models for Named Entity Recognition. In *Proceedings of CoNLL-2003, Edmonton, Canada, 2003.*

Chieu H. L., and Ng, H. T. (2003) Named Entity Recognition with a Maximum Entropy Approach. In *Proceedings of CoNLL-2003, Edmonton, Canada, 2003,* pp. 160-163.

Chincor, N. (1997). MUC-7 Named Entity Task Definition, Version 3.5.

Curran, J.R., and Clark, S. (2003). Language Independent NER using a Maximum Entropy Tagger. In *Proceedings of CoNLL-2003, Edmonton, Canada, 2003.*

Demiros, I., Boutsis, S., Giouli, V., Liakata, M., Papageorgiou, H., and Piperidis, S. (2000). Named Entity Recognition in Greek Texts. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation – LREC 2000, Athens, Greece, 31 May – 2 June 2000,* pp. 1223-1228.

Florian, R., Ittycheriah, A., Jing, H., and Zhang, T. (2003). Named Entity Recognition through Classifier Combination. In *Proceedings of CoNLL-2003, Edmonton, Canada, 2003.*

Klein, D., Smarz, J., Nguyen, H., and Manning, C. (2003). Named Entity Recognition with Character-Level Models. In *Proceedings of CoNLL-2003, Edmonton, Canada, 2003.*

Ng, H.T., Goh, W.B., and Low, K.L. (1997). Feature selection, Perceptron Learning and a Usability Case Study for Text Categorization. In *Proceedings of the 20th Annual Int. ACM SIGIR Conference on R&D in Information Retrieval (SIGIR*), pp 67-73.

Sang, E. F., and Kim, T. (2002). Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2002*, *Taipei, Taiwan, 2002*, pp. 155-158.

Sang, E. F.,Kim, T., and De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2003, Edmonton Canada, 2003,* pp. 142-147.