A survey of named entity recognition and classification

David Nadeau and Satoshi Sekine
National Research Council Canada / New York University

Introduction

The term "Named Entity", now widely used in Natural Language Processing, was coined for the Sixth Message Understanding Conference (MUC-6) (R. Grishman & Sundheim 1996). At that time, MUC was focusing on Information Extraction (IE) tasks where structured information of company activities and defense related activities is extracted from unstructured text, such as newspaper articles. In defining the task, people noticed that it is essential to recognize information units like names, including person, organization and location names, and numeric expressions including time, date, money and percent expressions. Identifying references to these entities in text was recognized as one of the important sub-tasks of IE and was called "Named Entity Recognition and Classification (NERC)".

We present here a survey of fifteen years of research in the NERC field, from 1991 to 2006. While early systems were making use of handcrafted rule-based algorithms, modern systems most often resort to machine learning techniques. We survey these techniques as well as other critical aspects of NERC such as features and evaluation methods. It was indeed concluded in a recent conference that the choice of features is at least as important as the choice of technique for obtaining a good NERC system (E. Tjong Kim Sang & De Meulder 2003). Moreover, the way NERC systems are evaluated and compared is essential to progress in the field. To the best of our knowledge, NERC features, techniques, and evaluation methods have not been surveyed extensively yet.

The first section of this survey presents some observations on published work from the point of view of activity per year, supported languages, preferred textual genre and domain, and supported entity types. It was collected from the review of a hundred English language papers sampled from the major conferences and journals. We do not claim this review to be exhaustive or representative of all the research in all languages, but we believe it gives a good feel for the breadth and depth of previous

Lingvisticæ Investigationes 30:1 (2007), **3–26**.
ISSN 0378–4169 / E-ISSN 1569–9927

work. Section 2 covers the algorithmic techniques that were proposed for addressing the NERC task. Most techniques are borrowed from the Machine Learning (ML) field. Instead of elaborating on techniques themselves, the third section lists and classifies the proposed features, i.e., descriptions and characteristic of words for algorithmic consumption. Section 4 presents some of the evaluation paradigms that were proposed throughout the major forums. Finally, we present our conclusions.

1. Observations: 1991 to 2006

The computational research aiming at automatically identifying named entities in texts forms a vast and heterogeneous pool of strategies, methods and representations. One of the first research papers in the field was presented by Lisa F. Rau (1991) at the Seventh IEEE Conference on Artificial Intelligence Applications. Rau's paper describes a system to "extract and recognize [company] names". It relies on heuristics and handcrafted rules. From 1991 (1 publication) to 1995 (we found 8 publications in English), the publication rate remained relatively low. It accelerated in 1996, with the first major event dedicated to the task: MUC-6 (R. Grishman & Sundheim 1996). It never declined since then with steady research and numerous scientific events: HUB-4 (N. Chinchor *et al.* 1998), MUC-7 and MET-2 (N. Chinchor 1999), IREX (S. Sekine & Isahara 2000), CONLL (E. Tjong Kim Sang 2002, E. Tjong Kim Sang & De Meulder 2003), ACE (G. Doddington *et al.* 2004) and HAREM (D. Santos *et al.* 2006). The Language Resources and Evaluation Conference (LREC)[1] has also been staging workshops and main conference tracks on the topic since 2000.

1.1 Language factor

A good proportion of work in NERC research is devoted to the study of English but a possibly larger proportion addresses language independence and multilingualism problems. German is well studied in CONLL-2003 and in earlier works. Similarly, Spanish and Dutch are strongly represented, boosted by a major devoted conference: CONLL-2002. Japanese has been studied in the MUC-6 conference, the IREX conference, and other works. Chinese is studied in an abundant literature (e.g., L.-J. Wang *et al.* 1992, H.-H. Chen & Lee 1996, S. Yu *et al.* 1998) and so are French (G. Petasis *et al.* 2001, Poibeau 2003), Greek (S. Boutsis *et al.* 2000), and Italian (W. Black *et al.* 1998, A. Cucchiarelli & Velardi 2001). Many other languages received some attention as well: Basque (C. Whitelaw & Patrick 2003), Bulgarian (J. Da Silva *et al.* 2004), Catalan (X. Carreras *et al.* 2003), Cebuano (J. May *et al.* 2003), Danish (E. Bick 2004), Hindi (S. Cucerzan & Yarowsky 1999, J. May *et al.* 2003), Korean (C. Whitelaw & Patrick 2003), Polish (J. Piskorski 2004), Romanian

(S. Cucerzan & Yarowsky 1999), Russian (B. Popov *et al*. 2004), Swedish (D. Kokkinakis 1998), and Turkish (S. Cucerzan & Yarowsky 1999). Portuguese was examined by (D. Palmer & Day 1997) and, at the time of writing this survey, the HAREM conference is revisiting that language. Finally, Arabic (F. Huang 2005) has started to receive a lot of attention in large-scale projects such as Global Autonomous Language Exploitation (GALE).[2]

1.2 Textual genre or domain factor

The impact of textual genre (journalistic, scientific, informal, etc.) and domain (gardening, sports, business, etc.) has been rather neglected in the NERC literature. Few studies are specifically devoted to diverse genres and domains. D. Maynard *et al*. (2001) designed a system for emails, scientific texts and religious texts. E. Minkov *et al.* (2005) created a system specifically designed for email documents. Perhaps unsurprisingly, these experiments demonstrated that although any domain can be reasonably supported, porting a system to a new domain or textual genre remains a major challenge. T. Poibeau and Kosseim (2001), for instance, tested some systems on both the MUC-6 collection composed of newswire texts, and on a proprietary corpus made of manual translations of phone conversations and technical emails. They report a drop in performance for every system (some 20% to 40% of precision and recall).

1.3 Entity type factor

In the expression "Named Entity", the word "Named" aims to restrict the task to only those entities for which one or many rigid designators, as defined by S. Kripke (1982), stands for the referent. For instance, *the automotive company created by Henry Ford in 1903* is referred to as *Ford* or *Ford Motor Company*. Rigid designators include proper names as well as certain natural kind terms like biological species and substances. There is a general agreement in the NERC community about the inclusion of temporal expressions and some numerical expressions such as amounts of money and other types of units. While some instances of these types are good examples of rigid designators (e.g., *the year 2001* is the 2001st year of the Gregorian calendar) there are also many invalid ones (e.g., *in June* refers to the month of an undefined year — *past June*, *this June*, *June 2020*, etc.). It is arguable that the NE definition is loosened in such cases for practical reasons.

Early work formulates the NERC problem as recognizing "proper names" in general (e.g., S. Coates-Stephens 1992, C. Thielen 1995). Overall, the most studied types are three specializations of "proper names": names of "persons", "locations" and "organizations". These types are collectively known as "enamex" since

the MUC-6 competition. The type "location" can in turn be divided into multiple subtypes of "fine-grained locations": city, state, country, etc. (M. Fleischman 2001, S. Lee & Geunbae Lee 2005). Similarly, "fine-grained person" sub-categories like "politician" and "entertainer" appear in the work of M. Fleischman and Hovy (2002). The type "person" is quite common and used at least once in an original way by O. Bodenreider and Zweigenbaum (2000) who combines it with other cues for extracting medication and disease names (e.g., "Parkinson disease"). In the ACE program, the type "facility" subsumes entities of the types "location" and "organization". The type "GPE" is used to represent a location which has a government, such as a city or a country.

The type "miscellaneous" is used in the CONLL conferences and includes proper names falling outside the classic "enamex". The class is also sometimes augmented with the type "product" (e.g., E. Bick 2004). The "timex" (another term coined in MUC) types "date" and "time" and the "numex" types "money" and "percent" are also quite predominant in the literature. Since 2003, a community named TIMEX2 (L. Ferro *et al.* 2005) proposes an elaborated standard for the annotation and normalization of temporal expressions. Finally, marginal types are sometime handled for specific needs: "film" and "scientist" (O. Etzioni *et al.* 2005), "email address" and "phone number" (I. Witten *et al.* 1999, D. Maynard *et al.* 2001), "research area" and "project name" (J. Zhu *et al.* 2005), "book title" (S. Brin 1998, I. Witten *et al.* 1999), "job title" (W. Cohen & Sarawagi 2004) and "brand" (E. Bick 2004).

A recent interest in bioinformatics, and the availability of the GENIA corpus (T. Ohta *et al.* 2002) led to many studies dedicated to types such as "protein", "DNA", "RNA", "cell line" and "cell type" (e.g., D. Shen *et al.* 2003, B. Settles 2004) as well as studies targeted at "protein" recognition only (Y. Tsuruoka & Tsujii 2003). Related work also includes "drug" (T. Rindfleisch *et al.* 2000) and "chemical" (M. Narayanaswamy *et al.* 2003) names.

Some recent work does not limit the possible types to extract and is referred to as "open domain" NERC (See E. Alfonseca & Manandhar 2002, R. Evans 2003). In this line of research, S. Sekine and Nobata (2004) defined a named entity hierarchy which includes many fine grained subcategories, such as museum, river or airport, and adds a wide range of categories, such as product and event, as well as substance, animal, religion, or color. It tries to cover most frequent name types and rigid designators appearing in a newspaper. The number of categories is about 200, and they are now defining popular attributes for each category to make it an ontology.

1.4 What's next?

Recent researches in multimedia indexing, semi-supervised learning, complex linguistic phenomena, and machine translation suggest some new directions for the

field. On one side, there is a growing interest in multimedia information processing (e.g., video, speech) and particularly NE extraction from it (R. Basili *et al.* 2005). Lot of effort is also invested toward semi-supervised and unsupervised approaches to NERC motivated by the use of very large collections of texts (O. Etzioni *et al.* 2005) and the possibility of handling multiple NE types (D. Nadeau *et al.* 2006). Complex linguistic phenomena (e.g., metonymy) that are common shortcomings of current systems are under investigation (T. Poibeau, 2006). Finally, large-scale projects such as GALE, discussed in Section 1.1, open the way to integration of NERC and Machine Translation for mutual improvement.

2. Learning methods

The ability to recognize previously unknown entities is an essential part of NERC systems. Such ability hinges upon recognition and classification rules triggered by distinctive features associated with positive and negative examples. While early studies were mostly based on handcrafted rules, most recent ones use supervised machine learning (SL) as a way to automatically induce rule-based systems or sequence labeling algorithms starting from a collection of training examples. This is evidenced, in the research community, by the fact that five systems out of eight were rule-based in the MUC-7 competition while sixteen systems were presented at CONLL-2003, a forum devoted to learning techniques. When training examples are not available, handcrafted rules remain the preferred technique, as shown in S. Sekine and Nobata (2004) who developed a NERC system for 200 entity types.

The idea of supervised learning is to study the features of positive and negative examples of NE over a large collection of annotated documents and design rules that capture instances of a given type. Section 2.1 explains SL approaches in more details. The main shortcoming of SL is the requirement of a large annotated corpus. The unavailability of such resources and the prohibitive cost of creating them lead to two alternative learning methods: semi-supervised learning (SSL) and unsupervised learning (UL). These techniques are presented in Section 2.2 and 2.3 respectively.

2.1 Supervised learning

The current dominant technique for addressing the NERC problem is supervised learning. SL techniques include Hidden Markov Models (HMM) (D. Bikel *et al.* 1997), Decision Trees (S. Sekine 1998), Maximum Entropy Models (ME) (A. Borthwick 1998), Support Vector Machines (SVM) (M. Asahara & Matsumoto 2003), and Conditional Random Fields (CRF) (A. McCallum & Li 2003). These are

all variants of the SL approach that typically consist of a system that reads a large annotated corpus, memorizes lists of entities, and creates disambiguation rules based on discriminative features.

A baseline SL method that is often proposed consists of tagging words of a test corpus when they are annotated as entities in the training corpus. The performance of the baseline system depends on the vocabulary transfer, which is the proportion of words, without repetition, appearing in both training and testing corpus. D. Palmer and Day (1997) calculated the vocabulary transfer on the MUC-6 training data. They report a transfer of 21%, with as much as 42% of location names being repeated but only 17% of organizations and 13% of person names. Vocabulary transfer is a good indicator of the recall (number of entities identified over the total number of entities) of the baseline system but is a pessimistic measure since some entities are frequently repeated in documents. A. Mikheev *et al.* (1999) precisely calculated the recall of the baseline system on the MUC-7 corpus. They report a recall of 76% for locations, 49% for organizations and 26% for persons with precision ranging from 70% to 90%. Whitelaw and Patrick (2003) report consistent results on MUC-7 for the aggregated enamex class. For the three enamex types together, the precision of recognition is 76% and the recall is 48%.

2.2 Semi-supervised learning

The term "semi-supervised" (or "weakly supervised") is relatively recent. The main technique for SSL is called "bootstrapping" and involves a small degree of supervision, such as a set of seeds, for starting the learning process. For example, a system aimed at "disease names" might ask the user to provide a small number of example names. Then the system searches for sentences that contain these names and tries to identify some contextual clues common to the five examples. Then, the system tries to find other instances of disease names that appear in similar contexts. The learning process is then reapplied to the newly found examples, so as to discover new relevant contexts. By repeating this process, a large number of disease names and a large number of contexts will eventually be gathered. Recent experiments in semi-supervised NERC (Nadeau *et al.* 2006) report performances that rival baseline supervised approaches. Here are some examples of SSL approaches.

S. Brin (1998) uses lexical features implemented by regular expressions in order to generate lists of book titles paired with book authors. It starts with seed examples such as {*Isaac Asimov, The Robots of Dawn*} and use some fixed lexical control rules such as the following regular expression *[A-Z][A-Za-z .,&]5,30[A-Z a-z.]* used to describe a title. The main idea of his algorithm, however, is that many web sites conform to a reasonably uniform format across the site. When a given web site is found to contain seed examples, new pairs can often be identified using

simple constraints such as the presence of identical text before, between or after the elements of an interesting pair. For example, the passage "*The Robots of Dawn, by Isaac Asimov (Paperback)*" would allow finding, on the same web site, "*The Ants, by Bernard Werber (Paperback)*".

M. Collins and Singer (1999) parse a complete corpus in search of candidate NE patterns. A pattern is, for instance, a proper name (as identified by a part-of-speech tagger) followed by a noun phrase in apposition (e.g., *Maury Cooper, a vice president at S&P*). Patterns are kept in pairs {*spelling, context*} where *spelling* refers to the proper name and *context* refers to the noun phrase in its context. Starting with an initial seed of spelling rules (e.g., *rule 1: if the spelling is "New York" then it is a Location; rule 2: if the spelling contains "Mr." then it is a Person; rule 3: if the spelling is all capitalized then it is an organization*), the candidates are examined. Candidates that satisfy a spelling rule are classified accordingly and their contexts are accumulated. The most frequent contexts found are turned into a set of contextual rules. Following the steps above, contextual rules can be used to find further spelling rules, and so on. M. Collins and Singer and R. Yangarber *et al.* (2002), demonstrate the idea that learning several types of NE simultaneously allows the finding of negative evidence (one type against all) and reduces over-generation. S. Cucerzan and Yarowsky (1999) also use a similar technique and apply it to many languages.

E. Riloff and Jones (1999) introduce mutual bootstrapping that consists of growing a set of entities and a set of contexts in turn. Instead of working with predefined candidate NEs (found using a fixed syntactic construct), they start with a handful of seed entity examples of a given type (e.g., *Bolivia*, *Guatemala*, *Honduras* are entities of type *country*) and accumulate all patterns found around these seeds in a large corpus. Contexts (e.g., *offices in X, facilities in X*, …) are ranked and used to find new examples. Riloff and Jones note that the performance of that algorithm can deteriorate rapidly when noise is introduced in the entity list or pattern list. While they report relatively low precision and recall in their experiments, their work proved to be highly influential.

A. Cucchiarelli and Velardi (2001) use syntactic relations (e.g., subject-object) to discover more accurate contextual evidence around the entities. Again, this is a variant of E. Riloff and Jones mutual bootstrapping (1999). Interestingly, instead of using human generated seeds, they rely on existing NER systems (called *early NE classifier*) for initial NE examples.

M. Pasca *et al.* (2006) are also using techniques inspired by mutual bootstrapping. However, they innovate through the use of D. Lin's (1998) distributional similarity to generate synonyms — or, more generally, words which are members of the same semantic class — allowing pattern generalization. For instance, for the pattern *X was born in November*, Lin's synonyms for *November* are {*March, October, April, Mar, Aug., February, Jul, Nov., …*} thus allowing the induction of new

patterns such as *X was born in March*. One of the contribution of Pasca *et al.* is to apply the technique to very large corpora (100 million web documents) and demonstrate that starting from a seed of 10 examples facts (defined as entities of type person paired with entities of type year — standing for the person's year of birth) it is possible to generate one million facts with a precision of about 88%.

The problem of unlabelled data selection is addressed by J. Heng and Grishman (2006). They show how an existing NE classifier can be improved using bootstrapping methods. The main lesson they report is that relying upon large collection of documents is not sufficient by itself. Selection of documents using information retrieval-like relevance measures and selection of specific contexts that are rich in proper names and coreferences bring the best results in their experiments.

2.3 Unsupervised learning

The typical approach in unsupervised learning is clustering. For example, one can try to gather named entities from clustered groups based on the similarity of context. There are other unsupervised methods too. Basically, the techniques rely on lexical resources (e.g., WordNet), on lexical patterns and on statistics computed on a large unannotated corpus. Here are some examples.

E. Alfonseca and Manandhar (2002) study the problem of labeling an input word with an appropriate NE type. NE types are taken from WordNet (e.g., location>country, animate>person, animate>animal, etc.). The approach is to assign a topic signature to each WordNet synset by merely listing words that frequently cooccur with it in a large corpus. Then, given an input word in a given document, the word context (words appearing in a fixed-size window around the input word) is compared to type signatures and classified under the most similar one.

In R. Evans (2003), the method for identification of hyponyms/hypernyms described in the work of M. Hearst (1992) is applied in order to identify potential hypernyms of sequences of capitalized words appearing in a document. For instance, when X is a capitalized sequence, the query "such as X", is searched on the web and, in the retrieved documents, the noun that immediately precedes the query can be chosen as the hypernym of X. Similarly, in P. Cimiano and Völker (2005), Hearst patterns are used but this time, the feature consists of counting the number of occurrences of passages like: "city such as", "organization such as", etc.

Y. Shinyama and Sekine (2004) used an observation that named entities often appear synchronously in several news articles, whereas common nouns do not. They found a strong correlation between being a named entity and appearing punctually (in time) and simultaneously in multiple news sources. This technique allows identifying rare named entities in an unsupervised manner and can be useful in combination with other NERC methods.

In O. Etzioni *et al.* (2005), Pointwise Mutual Information and Information Retrieval (PMI-IR) is used as a feature to assess that a named entity can be classified under a given type. PMI-IR, developed by P. Turney (2001), measures the dependence between two expressions using web queries. A high PMI-IR means that expressions tend to co-occur. O. Etzioni *et al.* create features for each candidate entity (e.g., London) and a large number of automatically generated discriminator phrases like "is a city", "nation of", etc.

3. Feature space for NERC

Features are descriptors or characteristic attributes of words designed for algorithmic consumption. An example of a feature is a Boolean variable with the value *true* if a word is capitalized and *false* otherwise. Feature vector representation is an abstraction over text where typically each word is represented by one or many Boolean, numeric and nominal values. For example, a hypothetical NERC system may represent each word in a text with 3 attributes:

1. a Boolean attribute with the value *true* if the word is capitalized and *false* otherwise;
2. a numeric attribute corresponding to the length, in characters, of the word;
3. a nominal attribute corresponding to the lowercased version of the word.

In this scenario, the sentence "The president of Apple eats an apple.", excluding the punctuation, would be represented by the following feature vectors:

<true, 3, "the">, <false, 9, "president">, <false, 2, "of">, <true, 5, "apple">, <false, 4, "eats">, <false, 2, "an">, <false, 5, "apple">

Usually, the NERC problem is resolved by applying a rule system over the features. For instance, a system might have two rules, a recognition rule: "capitalized words are candidate entities" and a classification rule: "the type of candidate entities of length greater than 3 words is organization". These rules work well for the exemplar sentence above. However, real systems tend to be much more complex and their rules are often created by automatic learning techniques.

In this section, we present the features most often used for the recognition and classification of named entities. We organize them along three different axes: word-level features, list lookup features and document and corpus features.

3.1 Word-level features

Word-level features are related to the character makeup of words. They specifically describe word case, punctuation, numerical value and special characters. Table 1 lists subcategories of word-level features.

Table 1. Word-level features

Features	Examples
Case	– Starts with a capital letter – Word is all uppercased – The word is mixed case (e.g., ProSys, eBay)
Punctuation	– Ends with period, has internal period (e.g., St., I.B.M.) – Internal apostrophe, hyphen or ampersand (e.g., O'Connor)
Digit	– Digit pattern (*see Section 3.1.1*) – Cardinal and ordinal – Roman number – Word with digits (e.g., W3C, 3M)
Character	– Possessive mark, first person pronoun – Greek letters
Morphology	– Prefix, suffix, singular version, stem – Common ending (*see Section 3.1.2*)
Part-of-speech	– proper name, verb, noun, foreign word
Function	– Alpha, non-alpha, n-gram (*see Section 3.1.3*) – lowercase, uppercase version – pattern, summarized pattern (*see Section 3.1.4*) – token length, phrase length

3.1.1 *Digit pattern*

Digits can express a wide range of useful information such as dates, percentages, intervals, identifiers, etc. Special attention must be given to some particular patterns of digits. For example, two-digit and four-digit numbers can stand for years (D. Bikel *et al.* 1997) and when followed by an "s", they can stand for a decade; one and two digits may stand for a day or a month (S. Yu *et al.* 1998).

3.1.2 *Common word ending*

Morphological features are essentially related to words affixes and roots. For instance, a system may learn that a human profession often ends in "ist" (e.g.,

journalist, *cyclist*) or that nationality and languages often ends in "ish" and "an" (e.g., *Spanish, Danish, Romanian*). Another example of common word ending is organization names that often end in "ex", "tech", and "soft" (E. Bick 2004).

3.1.3 *Functions over words*

Features can be extracted by applying functions over words. An example is given by M. Collins and Singer (1999) who create a feature by isolating the non-alphabetic characters of a word (e.g., nonalpha(A.T.&T.) = ..&.) Another example is given by J. Patrick *et al.* (2002) who use character n-grams as features.

3.1.4 *Patterns and summarized patterns*

Pattern features were introduced by M. Collins (2002) and then used by others (W. Cohen & Sarawagi 2004 and B. Settles 2004). Their role is to map words onto a small set of patterns over character types. For instance, a pattern feature might map all uppercase letters to "A", all lowercase letters to "a", all digits to "0" and all punctuation to "-":

x = "G.M.": GetPattern(x) = "A-A-"
x = "Machine-223": GetPattern(x) = "Aaaaaaa-000"

The summarized pattern feature is a condensed form of the above in which consecutive character types are not repeated in the mapped string. For instance, the preceding examples become:

x = "G.M.": GetSummarizedPattern(x) = "A-A-"
x = "Machine-223": GetSummarizedPattern(x) = "Aa-0"

3.2 List lookup features

Lists are the privileged features in NERC. The terms "gazetteer", "lexicon" and "dictionary" are often used interchangeably with the term "list". List inclusion is a way to express the relation "is a" (e.g., *Paris is a city*). It may appear obvious that if a word (*Paris*) is an element of a list of cities, then the probability of this word to be city, in a given text, is high. However, because of word polysemy, the probability is almost never 1 (e.g., the probability of "Fast" to represent a company is low because of the common adjective "fast" that is much more frequent).

In Table 2, we present three significant categories of lists used in literature. We could enumerate many more list examples but we decided to concentrate on those aimed at recognizing enamex types.

Table 2. List lookup features.

Features	Examples
General list	– General dictionary (see Section 3.2.1) – Stop words (function words) – Capitalized nouns (e.g., January, Monday) – Common abbreviations
List of entities	– Organization, government, airline, educational – First name, last name, celebrity – Astral body, continent, country, state, city
List of entity cues	– Typical words in organization (see 3.2.2) – Person title, name prefix, post-nominal letters – Location typical word, cardinal point

3.2.1 *General dictionary*

Common nouns listed in a dictionary are useful, for instance, in the disambiguation of capitalized words in ambiguous positions (e.g., sentence beginning). A. Mikheev (1999) reports that from 2677 words in ambiguous position in a given corpus, a general dictionary lookup allows identifying 1841 common nouns out of 1851 (99.4%) while only discarding 171 named entities out of 826 (20.7%). In other words, 20.7% of named entities are ambiguous with common nouns, in that corpus.

3.2.2 *Words that are typical of organization names*

Many authors propose to recognize organizations by identifying words that are frequently used in their names. For instance, knowing that "associates" is frequently used in organization names could lead to the recognition of "Computer Associates" and "BioMedia Associates" (D. McDonald 1993, R. Gaizauskas *et al.* 1995). The same rule applies to frequent first words ("American", "General") of an organization (L. Rau 1991). Some authors also exploit the fact that organizations often include the name of a person (F. Wolinski *et al.* 1995, Y. Ravin & Wacholder 1996) as in "Alfred P. Sloan Foundation". Similarly, geographic names can be good indicators of an organization name (F. Wolinski *et al.* 1995) as in "France Telecom". Organization designators such as "inc" and "corp" (L. Rau 1991) are also useful features.

3.2.3 *On the list lookup techniques*

Most approaches implicitly require candidate words to exactly match at least one element of a pre-existing list. However, we may want to allow some flexibility in the match conditions. At least three alternate lookup strategies are used in the NERC field.

First, words can be stemmed (stripping off both inflectional and derivational suffixes) or lemmatized (normalizing for inflections only) before they are matched (S. Coates-Stephens 1992). For instance, if a list of cue words contains "technology", the inflected form "technologies" will be considered as a successful match. For some languages (M. Jansche 2002), diacritics can be replaced by their canonical equivalent (e.g., 'é' replaced by 'e').

Second, candidate words can be "fuzzy-matched" against the reference list using some kind of thresholded edit-distance (Y. Tsuruoka & Tsujii 2003) or Jaro-Winkler (W. Cohen & Sarawagi 2004). This allows capturing small lexical variations in words that are not necessarily derivational or inflectional. For instance, *Frederick* could match *Frederik* because the edit-distance between the two words is very small (suppression of just one character, the 'c'). Jaro-Winkler's metric was specifically designed to match proper names following the observation that the first letters tend to be correct while name ending often varies.

Third, the reference list can be accessed using the Soundex algorithm (H. Raghavan & Allan 2004) which normalizes candidate words to their respective Soundex codes. This code is a combination of the first letter of a word plus a three digit code that represents its phonetic sound. Hence, similar sounding names like *Lewinskey* (soundex = *l520*) and *Lewinsky* (soundex = *l520*) are equivalent with respect to their Soundex code.

3.3 Document and corpus features

Document features are defined by both document content and document structure. Large collections of documents (corpora) are also excellent sources of features. We list in this section features that go beyond the single word and multi-

Table 3. Features from documents.

Features	Examples
Multiple occurrences	– Other entities in the context – Uppercased and lowercased occurrences (see 3.3.1) – Anaphora, coreference (see 3.3.2)
Local syntax	– Enumeration, apposition – Position in sentence, in paragraph, and in document
Meta information	– Uri, email header, XML section, (see Section 3.3.3) – Bulleted/numbered lists, tables, figures
Corpus frequency	– Word and phrase frequency – Co-occurrences – Multiword unit permanency (see 3.3.4)

word expression and include meta-information about documents and corpus statistics.

3.3.1 *Multiple occurrences and multiple casing*

C. Thielen (1995), Y. Ravin and Wacholder (1996) and A. Mikheev (1999) identify words that appear both in uppercased and lowercased form in a single document. These words are hypothesized to be common nouns that appear both in ambiguous (e.g., sentence beginning) and unambiguous position.

3.3.2 *Entity coreference and alias*

The task of recognizing the multiple occurrences of a unique entity in a document dates back to the earliest research in the field (D. McDonald 1993, L. Rau 1991). Coreferences are the occurrences of a given word or word sequence referring to a given entity within a document. Deriving features from coreferences is mainly done by exploiting the context of every occurrence (e.g., *Macdonald was the first*, *Macdonald said*, *was signed by Macdonald*, …). Aliases of an entity are the various ways the entity is written in a document. For instance, we may have the following aliases for a given entity: *Sir John A. Macdonald*, *John A. Macdonald*, *John Alexander Macdonald*, and *Macdonald*. Deriving features from aliases is mainly done by leveraging the union of alias words (*Sir*, *John*, *A*, *Alexander*, *Macdonald*).

Finding coreferences and aliases in a text can be reduced to the same problem of finding all occurrences of an entity in a document. This problem is of great complexity. R. Gaizauskas *et al.* (1995) use 31 heuristic rules to match multiple occurrences of company names. For instance, two multi-word expressions match if one is the initial subsequence of the other. An even more complex task is the recognition of entity mention across documents. X. Li *et al.* (2004) propose and compare a supervised and an unsupervised model for this task. They propose the use of word-level features engineered to handle equivalences (e.g., *prof.* is equivalent to *professor*) and relational features to encode the relative order of tokens between two occurrences.

Word-level features are often insufficient for complex problems. A metonymy, for instance, denotates a different concept than the literal denotation of a word (e.g., "New York" that stands for "New York Yankees", "Hexagon" that stands for "France"). T. Poibeau (2006) shows that semantic tagging is a key issue in such case.

3.3.3 *Document meta-information*

Most meta-information about documents can be used directly: email headers are good indicators of person names, news often starts with a location name, etc. Some authors make original use of meta-information. J. Zhu *et al.* (2005) uses document

URL to bias probabilities of entities. For instance, many names (e.g., bird names) have high probability to be a "project name" if the URL is from a computer science department domain.

3.3.4 *Statistics for multiword units*

J. Da Silva *et al.* (2004) propose some interesting feature functions for multi-word units that can be thresholded using corpus statistics. For example, they establish a threshold on the presence of rare and long lowercased words in entities. Only multiword units that do not contain rare lowercased words (rarity calculated as relative frequency in the corpus) of a relatively long size (mean size calculated from the corpus) are considered as candidate named entities. They also present a feature called permanency that consists of calculating the frequency of a word (e.g., *Life*) in a corpus divided by its frequency in case insensitive form (e.g., *life*, *Life*, *LIFE*, etc.)

4. Evaluation of NERC

Thorough evaluation of NERC systems is essential to their progress. Many techniques were proposed to rank systems based on their capability to annotate a text like an expert linguist. In the following section, we take a look at three main scoring techniques used for MUC, IREX, CONLL and ACE conferences. But first, let's summarize the task from the point of view of evaluation.

In NERC, systems are usually evaluated based on how their output compares with the output of human linguists. For instance, here's an annotated text marked up according to the MUC guidelines. Let's call it the solution.

> Unlike <ENAMEX TYPE="PERSON">Robert</ENAMEX>, <ENAMEX TYPE="PERSON">John Briggs Jr</ENAMEX> contacted <ENAMEX TYPE="ORGANIZATION">Wonderful Stockbrockers Inc</ENAMEX> in <ENAMEX TYPE="LOCATION">New York</ENAMEX> and instructed them to sell all his shares in <ENAMEX TYPE="ORGANIZATION">Acme </ENAMEX>.

Let's now hypothesize a system producing the following output:

> <ENAMEX TYPE="LOCATION">Unlike</ENAMEX> Robert, <ENAMEX TYPE="ORGANIZATION">John Briggs Jr</ENAMEX> contacted Wonderful <ENAMEX TYPE="ORGANIZATION">Stockbrockers</ ENAMEX> Inc <ENAMEX TYPE="PERSON">in New York</ENAMEX> and instructed them to sell all his shares in <ENAMEX TYPE="ORGANIZA TION">Acme</ENAMEX>.

The system produced five different errors,[3] explained in Table 4. In this example, the system gives one correct answer: (<Organization> Acme </Organization>). Ultimately, the question is "What score should we give to this system?" In the following sections, we survey how the question was answered in various evaluation forums.

Table 4. NERC error types.

Correct solution	System output	Error
Unlike	<ENAMEX TYPE="LOCATION"> Unlike </ENAMEX>	The system hypothesized an entity where there is none.
<ENAMEX TYPE="PERSON"> Robert </ENAMEX>	Robert	An entity was completely missed by the system.
<ENAMEX TYPE="PERSON"> John Briggs Jr </ENAMEX>	<ENAMEX TYPE="ORGANIZATION"> John Briggs Jr </ENAMEX>	The system noticed an entity but gave it the wrong label.
<ENAMEX TYPE="ORGANIZATION"> Wonderful Stockbrockers Inc </ENAMEX>	<ENAMEX TYPE="ORGANIZATION"> Stockbrockers </ENAMEX>	A system noticed there is an entity but got its boundaries wrong.
<ENAMEX TYPE="LOCATION"> New York </ENAMEX>	<ENAMEX TYPE="PERSON"> in New York </ENAMEX>	The system gave the wrong label to the entity and got its boundary wrong.

4.1 MUC evaluations

In MUC events (R. Grishman & Sundheim 1996, N. Chinchor 1999), a system is scored on two axes: its ability to find the correct type (TYPE) and its ability to find exact text (TEXT). A correct TYPE is credited if an entity is assigned the correct type, regardless of boundaries as long as there is an overlap. A correct TEXT is credited if entity boundaries are correct, regardless of the type. For both TYPE and TEXT, three measures are kept: the number of correct answers (COR), the number of actual system guesses (ACT) and the number of possible entities in the solution (POS).

The final MUC score is the micro-averaged f-measure (MAF), which is the harmonic mean of precision and recall calculated over all entity slots on both axes. A micro-averaged measure is performed on all entity types without distinction (errors and successes for all entity types are summed together). The harmonic

mean of two numbers is never higher than the geometrical mean. It also tends toward the least number, minimizing the impact of large outliers and maximizing the impact of small ones. The F-measure therefore tends to privilege balanced systems.

In MUC, precision is calculated as COR / ACT and the recall is COR / POS. For the previous example, COR = 4 (2 TYPE + 2 TEXT), ACT = 10 (5 TYPE + 5 TEXT) and POS = 10 (5 TYPE + 5 TEXT). The precision is therefore 40%, the recall is 40% and the MAF is 40%.

This measure has the advantage of taking into account all possible types of errors of Table 4. It also gives partial credit for errors occurring on one axis only. Since there are two evaluation axes, each complete success is worth two points. The worst errors cost this two points (missing both TYPE and TEXT) while other errors cost only one point.

4.2 Exact-match evaluations

IREX and CONLL share a simple scoring protocol. We can call it "exact-match evaluation". Systems are compared based on the micro-averaged f-measure (MAF) with the precision being the percentage of named entities found by the system that are correct and the recall being the percentage of named entities present in the solution that are found by the system. A named entity is correct only if it is an exact match with the corresponding entity in the solution.

For the previous example, there are 5 true entities, 5 system guesses and only one guess that exactly matches the solution. The precision is therefore 20%, the recall is 20% and the MAF is 20%.

For some applications, the constraint of exact match is unnecessarily stringent. For instance, in some bioinformatics work, the goal is to determine whether or not a particular sentence mentions a specific gene and its function. Exact NE boundaries are not required: all is needed is to determine if the sentence does refer to the gene (R. Tzong-Han Tsai *et al.* 2006).

4.3 ACE evaluation

ACE has a complex evaluation procedure. It includes mechanisms for dealing various evaluation issues (partial match, wrong type, etc.). The ACE task definition is also more elaborated than previous tasks at the level of named entity "subtypes", "class" as well as entity mentions (coreferences), and more, but these supplemental elements will be ignored here.

Basically, each entity type has a parameterized weight and contributes up to a maximal proportion (MAXVAL) of the final score (e.g., if each person is worth

1 point and each organization is worth 0.5 point then it takes two organizations to counterbalance one person in the final score). Some entity types such as "facility" may account for as little as 0.05 points, according to ACE parameters. In addition, customizable costs (COST) are used for false alarms, missed entities and type errors. Partial matches of textual spans are only allowed if named entity head matches on at least a given proportion of characters. Temporal expressions are not treated in ACE since they are evaluated by the TIMEX2 community (L. Ferro *et al.* 2005).

The final score called Entity Detection and Recognition Value (EDR) is 100% minus the penalties. For the examples of Table 4, the EDR score is 31.3%. It is computed as follows, using ACE parameters from 2004.[4] Each of the five entities contributes up to a maximum value to the final score. Using default ACE parameters, the maximal values (MAXVAL) for person entities is 61.54% of the final score, the two organizations worth 30.77% and the location worth 7.69%. These values sum up to 100%. At the individual type level, one person span was recognized (*John Briggs Jr*) but with the wrong type (*organization*); one person entity was missed (*Robert*); the two organization spans (*Wonderful Stockbrockers Inc* and *Acme*) were considered correct, even if the former partially matches; one geopolitical span was recognized (*in New York*) but with the wrong type and there was one false alarm (*Unlike*). Globally, the error (function of COST and MAXVAL) for the person entities accounts for 55.31% of the final EDR loss (30.77 for the miss and 24.54 for the type error), the false alarm account for 5.77% of loss and the location type error accounts for 7.58%. The final EDR of 31.3% is 100% minus these losses.

ACE evaluation may be the most powerful evaluation scheme because of its customizable cost of error and its wide coverage of the problem. It is however problematic because the final scores are only comparable when parameters are fixed. In addition, complex methods are not intuitive and make error analysis difficult.

5. Conclusion

The Named Entity Recognition field has been thriving for more than fifteen years. It aims at extracting and classifying mentions of rigid designators, from text, such as proper names, biological species, and temporal expressions. In this survey, we have shown the diversity of languages, domains, textual genres and entity types covered in the literature. More than twenty languages and a wide range of named entity types are studied. However, most of the work has concentrated on limited domains and textual genres such as news articles and web pages.

We have also provided an overview of the techniques employed to develop NERC systems, documenting the recent trend away from hand-crafted rules

towards machine learning approaches. Handcrafted systems provide good performance at a relatively high system engineering cost. When supervised learning is used, a prerequisite is the availability of a large collection of annotated data. Such collections are available from the evaluation forums but remain rather rare and limited in domain and language coverage. Recent studies in the field have explored semi-supervised and unsupervised learning techniques that promise fast deployment for many entity types without the prerequisite of an annotated corpus. We have listed and categorized the features that are used in recognition and classification algorithms. The use of an expressive and varied set of features turns out to be just as important as the choice of machine learning algorithms. Finally we have also provided an overview of the evaluation methods that are in use in the major forums of the NERC research community. We saw that in a simple example made of only five named entities, the score of three different evaluation techniques vary from 20% to 40%.

NERC will have a profound impact on our society. Early commercial initiatives are already modifying the way we use yellow pages by providing local search engines (search your neighborhood for organizations, product and services, people, etc.). NERC systems also enable monitoring trends in the huge space of textual media produced every day by organizations, governments and individuals. It is also at the basis of a major advance in biology and genetics, enabling researchers to search the abundant literature for interactions between named genes and cells.

Acknowledgement

Thanks to Ralph Grishman, Stan Matwin and Peter D. Turney for helpful comments.

Notes

1. http://www.lrec-conf.org/

2. http://projects.ldc.upenn.edu/gale/

3. Types of errors are inspired by an informal publication by Christopher Manning: http://nlpers.blogspot.com/2006/08/doing-named-entity-recognition-dont.html

4. http://www.nist.gov/speech/tests/ace/ace04/index.htm

References

Alfonseca, Enrique; Manandhar, S. 2002. An Unsupervised Method for General Named Entity Recognition and Automated Concept Discovery. In *Proc. International Conference on General WordNet*.

Asahara, Masayuki; Matsumoto, Y. 2003. Japanese Named Entity Extraction with Redundant Morphological Analysis. In *Proc. Human Language Technology conference — North American chapter of the Association for Computational Linguistics*.

Basili, Roberto; Cammisa, M.; Donati, E. 2005. RitroveRAI: A Web Application for Semantic Indexing and Hyperlinking of Multimedia News. In *Proc. International Semantic Web Conference*.

Bick, Eckhard. 2004. A Named Entity Recognizer for Danish. In *Proc. Conference on Language Resources and Evaluation*.

Bikel, Daniel M.; Miller, S.; Schwartz, R.; Weischedel, R. 1997. Nymble: a High-Performance Learning Name-finder. In *Proc. Conference on Applied Natural Language Processing*.

Black, William J.; Rinaldi, F.; Mowatt, D. 1998. Facile: Description of the NE System used for Muc-7. In *Proc. Message Understanding Conference*.

Bodenreider, Olivier; Zweigenbaum, P. 2000. Identifying Proper Names in Parallel Medical Terminologies. *Stud Health Technol Inform* 77.443–447, Amsterdam: IOS Press.

Boutsis, Sotiris; Demiros, I.; Giouli, V.; Liakata, M.; Papageorgiou, H.; Piperidis, S. 2000. A System for Recognition of Named Entities in Greek. In *Proc. International Conference on Natural Language Processing*.

Borthwick, Andrew; Sterling, J.; Agichtein, E.; Grishman, R. 1998. NYU: Description of the MENE Named Entity System as used in MUC-7. In *Proc. Seventh Message Understanding Conference*.

Brin, Sergey. 1998. Extracting Patterns and Relations from the World Wide Web. In *Proc. Conference of Extending Database Technology. Workshop on the Web and Databases*.

Carreras, Xavier; Márques, L.; Padró, L. 2003. Named Entity Recognition for Catalan Using Spanish Resources. In *Proc. Conference of the European Chapter of Association for Computational Linguistic*.

Chen, H. H.; Lee, J. C. 1996. Identification and Classification of Proper Nouns in Chinese Texts. In *Proc. International Conference on Computational Linguistics*.

Chinchor, Nancy. 1999. Overview of MUC-7/MET-2. In *Proc. Message Understanding Conference MUC-7*.

Chinchor, Nancy; Robinson, P.; Brown, E. 1998. Hub-4 Named Entity Task Definition. In *Proc. DARPA Broadcast News Workshop*.

Cimiano, Philipp; Völker, J. 2005. Towards Large-Scale, Open-Domain and Ontology-Based Named Entity Classification. In *Proc. Conference on Recent Advances in Natural Language Processing*.

Coates-Stephens, Sam. 1992. The Analysis and Acquisition of Proper Names for the Understanding of Free Text. *Computers and the Humanities* 26.441–456, San Francisco: Morgan Kaufmann Publishers.

Cohen, William W.; Sarawagi, S. 2004. Exploiting Dictionaries in Named Entity Extraction: Combining Semi-Markov Extraction Processes and Data Integration Methods. In *Proc. Conference on Knowledge Discovery in Data*.

Collins, Michael. 2002. Ranking Algorithms for Named-Entity Extraction: Boosting and the Voted Perceptron. In *Proc. Association for Computational Linguistics*.

Collins, Michael; Singer, Y. 1999. Unsupervised Models for Named Entity Classification. In *Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Cucchiarelli, Alessandro; Velardi, P. 2001. Unsupervised Named Entity Recognition Using Syntactic and Semantic Contextual Evidence. *Computational Linguistics* 27:1.123–131, Cambridge: MIT Press.

Cucerzan, Silviu; Yarowsky, D. 1999. Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. In *Proc. Joint Sigdat Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Da Silva, Joaquim Ferreira; Kozareva, Z.; Lopes, G. P. 2004. Cluster Analysis and Classification of Named Entities. In *Proc. Conference on Language Resources and Evaluation*.

Doddington, George; Mitchell, A.; Przybocki, M.; Ramshaw, L.; Strassel, S.; Weischedel, R. 2004. The Automatic Content Extraction (ACE) Program — Tasks, Data, and Evaluation. In *Proc. Conference on Language Resources and Evaluation*.

Etzioni, Oren; Cafarella, M.; Downey, D.; Popescu, A.-M.; Shaked, T.; Soderland, S.; Weld, D. S.; Yates, A. 2005. Unsupervised Named-Entity Extraction from the Web: An Experimental Study. *Artificial Intelligence* 165.91–134, Essex: Elsevier Science Publishers.

Evans, Richard. 2003. A Framework for Named Entity Recognition in the Open Domain. In *Proc. Recent Advances in Natural Language Processing*.

Ferro, Lisa; Gerber, L.; Mani, I.; Sundheim, B.; Wilson G. 2005. *TIDES 2005 Standard for the Annotation of Temporal Expressions*. The MITRE Corporation.

Fleischman, Michael. 2001. Automated Subcategorization of Named Entities. In *Proc. Conference of the European Chapter of Association for Computational Linguistic*.

Fleischman, Michael; Hovy. E. 2002. Fine Grained Classification of Named Entities. In *Proc. Conference on Computational Linguistics*.

Gaizauskas, Robert.; Wakao, T.; Humphreys, K.; Cunningham, H.; Wilks, Y. 1995. University of Sheffield: Description of the LaSIE System as Used for MUC-6. In *Proc. Message Understanding Conference*.

Grishman, Ralph; Sundheim, B. 1996. Message Understanding Conference-6: A Brief History. In *Proc. International Conference on Computational Linguistics*.

Hearst, Marti. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proc. International Conference on Computational Linguistics*.

Heng, Ji; Grishman, R. 2006. Data Selection in Semi-supervised Learning for Name Tagging. In *Proc. joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics. Information Extraction beyond the Document.*

Huang, Fei. 2005. *Multilingual Named Entity Extraction and Translation from Text and Speech*. Ph.D. Thesis. Pittsburgh: Carnegie Mellon University.

Jansche, Martin. 2002. Named Entity Extraction with Conditional Markov Models and Classifiers. In *Proc. Conference on Computational Natural Language Learning*.

Kokkinakis, Dimitri. 1998., AVENTINUS, GATE and Swedish Lingware. In *Proc. of Nordic Computational Linguistics Conference*.

Kripke, Saul. 1982. *Naming and Necessity*. Boston: Harvard University Press.

Lee, Seungwoo; Geunbae Lee, G. 2005. Heuristic Methods for Reducing Errors of Geographic Named Entities Learned by Bootstrapping. In *Proc. International Joint Conference on Natural Language Processing*.

Li, Xin.; Morie, P.; Roth, D. 2004. Identification and Tracing of Ambiguous Names: Discriminative and Generative Approaches. In *Proc. National Conference on Artificial Intelligence.*

Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. In *Proc. International Conference on Computational Linguistics and the Annual Meeting of the Association for Computational Linguistics*.

McDonald, David D. 1993. Internal and External Evidence in the Identification and Semantic Categorization of Proper Names. In *Proc. Corpus Processing for Lexical Acquisition*.

May, Jonathan; Brunstein, A.; Natarajan, P.; Weischedel, R. M. 2003. Surprise! What's in a Cebuano or Hindi Name? *ACM Transactions on Asian Language Information Processing* 2:3.169–180, New York: ACM Press.

Maynard, Diana; Tablan, V.; Ursu, C.; Cunningham, H.; Wilks, Y. 2001. Named Entity Recognition from Diverse Text Types. In *Proc. Recent Advances in Natural Language Processing.*

McCallum, Andrew; Li, W. 2003. Early Results for Named Entity Recognition with Conditional Random Fields, Features Induction and Web-Enhanced Lexicons. In *Proc. Conference on Computational Natural Language Learning.*

Mikheev, Andrei. 1999. A Knowledge-free Method for Capitalized Word Disambiguation. In *Proc. Conference of Association for Computational Linguistics*.

Mikheev, A.; Moens, M.; Grover, C. 1999. Named Entity Recognition without Gazetteers. In *Proc. Conference of European Chapter of the Association for Computational Linguistics*.

Minkov, Einat; Wang, R.; Cohen, W. 2005. Extracting Personal Names from Email: Applying Named Entity Recognition to Informal Text. In *Proc. Human Language Technology and Conference Conference on Empirical Methods in Natural Language Processing.*

Nadeau, David; Turney, P.; Matwin, S. 2006. Unsupervised Named Entity Recognition: Generating Gazetteers and Resolving Ambiguity. In *Proc. Canadian Conference on Artificial Intelligence*.

Narayanaswamy, Meenakshi; Ravikumar K. E.; Vijay-Shanker K. 2003. A Biological Named Entity Recognizer. In *Proc. Pacific Symposium on Biocomputing.*

Ohta, Tomoko; Tateisi, Y.; Kim, J.; Mima, H.; Tsujii, J. 2002. The GENIA Corpus: An Annotated Research Abstract Corpus in Molecular Biology Domain. In *Proc. Human Language Technology Conference.*

Pasca, Marius; Lin, D.; Bigham, J.; Lifchits, A.; Jain, A. 2006. Organizing and Searching the World Wide Web of Facts — Step One: The One-Million Fact Extraction Challenge. In *Proc. National Conference on Artificial Intelligence.*

Patrick, Jon; Whitelaw, C.; Munro, R. 2002. SLINERC: The Sydney Language-Independent Named Entity Recogniser and Classifier. In *Proc. Conference on Natural Language Learning*.

Palmer, David D.; Day, D. S. 1997. A Statistical Profile of the Named Entity Task. In *Proc. ACL Conference for Applied Natural Language Processing.*

Petasis, Georgios; Vichot, F.; Wolinski, F.; Paliouras, G.; Karkaletsis, V.; Spyropoulos, C. D. 2001. Using Machine Learning to Maintain Rule-based Named-Entity Recognition and Classification Systems. In *Proc. Conference of Association for Computational Linguistics*.

Piskorski, Jakub. 2004. Extraction of Polish Named-Entities. In *Proc. Conference on Language Resources an Evaluation*.

Poibeau, Thierry. 2003. The Multilingual Named Entity Recognition Framework. In *Proc. Conference on European chapter of the Association for Computational Linguistics.*

Poibeau, Thierry. 2006. Dealing with Metonymic Readings of Named Entities. In *Proc. Annual Conference of the Cognitive Science Society*.

Poibeau, Thierry; Kosseim, L. 2001. Proper Name Extraction from Non-Journalistic Texts. In *Proc. Computational Linguistics in the Netherlands*.

Popov, Borislav; Kirilov, A.; Maynard, D.; Manov, D. 2004. Creation of reusable components and language resources for Named Entity Recognition in Russian. In *Proc. Conference on Language Resources and Evaluation*.

Raghavan, Hema; Allan, J. 2004. Using Soundex Codes for Indexing Names in ASR documents. In *Proc. Human Language Technology conference — North American chapter of the Association for Computational Linguistics. Interdisciplinary Approaches to Speech Indexing and Retrieval*.

Rau, Lisa F. 1991. Extracting Company Names from Text. In *Proc. Conference on Artificial Intelligence Applications of IEEE*.

Ravin, Yael; Wacholder, N. 1996. *Extracting Names from Natural-Language Text*. IBM Research Report RC 2033.

Riloff, Ellen; Jones, R 1999. Learning Dictionaries for Information Extraction using Multi-level Bootstrapping. In *Proc. National Conference on Artificial Intelligence*.

Rindfleisch, Thomas C.; Tanabe, L.; Weinstein, J. N. 2000. EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature. In *Proc. Pacific Symposium on Biocomputing*.

Santos, Diana; Seco, N.; Cardoso, N.; Vilela, R. 2006. HAREM: An Advanced NER Evaluation Contest for Portuguese. In *Proc. International Conference on Language Resources and Evaluation*.

Sekine, Satoshi. 1998. Nyu: Description of the Japanese NE System Used For Met-2. In *Proc. Message Understanding Conference*.

Sekine, Satoshi; Isahara, H. 2000. IREX: IR and IE Evaluation project in Japanese. In *Proc. Conference on Language Resources and Evaluation*.

Sekine, Satoshi; Nobata, C. 2004. Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. In *Proc. Conference on Language Resources and Evaluation*.

Settles, Burr. 2004. Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets. In *Proc. Conference on Computational Linguistics. Joint Workshop on Natural Language Processing in Biomedicine and its Applications*.

Shen Dan; Zhang, J.; Zhou, G.; Su, J.; Tan, C. L. 2003. Effective Adaptation of a Hidden Markov Model-based Named Entity Recognizer for Biomedical Domain. In *Proc. Conference of Association for Computational Linguistics. Natural Language Processing in Biomedicine*.

Shinyama, Yusuke; Sekine, S. 2004. Named Entity Discovery Using Comparable News Articles. In *Proc. International Conference on Computational Linguistics*.

Thielen, Christine. 1995. An Approach to Proper Name Tagging for German. In *Proc. Conference of European Chapter of the Association for Computational Linguistics. SIGDAT*.

Tjong Kim Sang, Erik. F. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *Proc. Conference on Natural Language Learning*.

Tjong Kim Sang, Erik. F.; De Meulder, F. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proc. Conference on Natural Language Learning*.

Tsuruoka, Yoshimasa; Tsujii, J. 2003. Boosting Precision and Recall of Dictionary-Based Protein Name Recognition. In *Proc. Conference of Association for Computational Linguistics. Natural Language Processing in Biomedicine*.

Turney, Peter. 2001. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In *Proc. European Conference on Machine Learning*.

Tzong-Han Tsai, Richard; Wu S.-H.; Chou, W.-C.; Lin, Y.-C.; He, D.; Hsiang, J.; Sung, T.-Y.; Hsu, W.-L. 2006. Various Criteria in the Evaluation of Biomedical Named Entity Recognition. *BMC Bioinformatics* 7:92, BioMed Central.

Wang, Liang-Jyh; Li, W.-C.; Chang, C.-H. 1992. Recognizing Unregistered Names for Mandarin Word Identification. In *Proc. International Conference on Computational Linguistics*.

Whitelaw, Casey; Patrick, J. 2003. Evaluating Corpora for Named Entity Recognition Using Character-Level Features. In *Proc. Australian Conference on Artificial Intelligence*.

Witten, Ian. H.; Bray, Z.; Mahoui, M.; Teahan W. J. 1999. Using Language Models for Generic Entity Extraction. In *Proc. International Conference on Machine Learning. Text Mining*.

Wolinski, Francis; Vichot, F.; Dillet, B. 1995. Automatic Processing Proper Names in Texts. In *Proc. Conference on European Chapter of the Association for Computational Linguistics*.

Yangarber, Roman; Lin, W.; Grishman, R. 2002. Unsupervised Learning of Generalized Names. In *Proc. of International Conference on Computational Linguistics.*

Yu, Shihong; Bai S.; Wu, P. 1998. Description of the Kent Ridge Digital Labs System Used for MUC-7. In *Proc. Message Understanding Conference*.

Zhu, Jianhan; Uren, V.; Motta, E. 2005. ESpotter: Adaptive Named Entity Recognition for Web Browsing. In *Proc. Conference Professional Knowledge Management. Intelligent IT Tools for Knowledge Management Systems*.

Summary

This survey covers fifteen years of research in the Named Entity Recognition and Classification (NERC) field, from 1991 to 2006. We report observations about languages, named entity types, domains and textual genres studied in the literature. From the start, NERC systems have been developed using hand-made rules, but now machine learning techniques are widely used. These techniques are surveyed along with other critical aspects of NERC such as features and evaluation methods. Features are word-level, dictionary-level and corpus-level representations of words in a document. Evaluation techniques, ranging from intuitive exact match to very complex matching techniques with adjustable cost of errors, are an indisputable key to progress.

Keywords: named identity, survey, learning method, feature space, evaluation.

Authors' addresses:

David Nadeau
National Research Council Canada
101 St-Jean-Bosco Street
Gatineau, QC, K1A 0R6
Canada

David.Nadeau@cnrc-nrc.gc.ca

Satoshi Sekine
New York University
715 Broadway, 7th floor
New York, NY 10003
USA

sekine@cs.nyu.edu