HÁSKÓLINN Í REYKJAVÍK
REYKJAVÍK UNIVERSITY

Named Entity Recognition for Icelandic
Research Report

Aðalsteinn Tryggvason

Leiðbeinandi: Hrafn Loftsson
Prófdómari: Hannes Högni Vilhjálmsson

Tölvunarfræðideild
Vor 2009

Named Entity Recognition for Icelandic

Abstract

I present an Icelandic Named Entity Recognition (NER) system. The goal of NER, which is a subtask of information extraction, is to locate and classify proper names, such as company names, person names, locations etc. As there is no Icelandic corpus available with manually tagged named entities, this system is developed using linguistic rules rather than statistical data. The system uses components (such as IceTagger) from the Natural Language Processing Toolkit for Icelandic (IceNLP) (Loftsson, H. and Rögnvaldsson, E. 2007) and will, when finished, become part of it.

Þessi skýrsla fjallar um nafnaþekkjara fyrir íslensku. Tilgangurinn með nafnaþekkjaranum er að finna og flokka sérnöfn, svo sem, nöfn fyrirtækja og stofnana, mannanöfn, örnefni o.s.frv. Þar sem ekki er til íslensk málheild þar sem búið er að handmerkja nöfn, þá verður nafnaþekkjarinn útfærður með málfræðilegum aðferðum fremur en aðferðum byggðum á tölfræði. Kerfið styðst við máltæknieingar (s.s. IceTagger) úr Natural Language Processing Toolkit for Icelandic (IceNLP) (Loftsson, H. and Rögnvaldsson, E. 2007) og verður að lokum hluti af því safni.

1. Introduction

Natural language processing (NLP) is a field of computer science that deals with interactions between computers and human languages. One of its subfields is information extraction (IE) whose goal is to extract structured information from text. This information can be, for example, names of places, persons or companies and in that case the subfield is known as *named entity recognition (NER)*. As more and more texts becomes available in computer readable format every day, methods to extract information from it become more vital.

There are various issues that have to be addressed when recognizing names. Where does the name start, where does it end and what kind of an entity is it: company, person or location? Well, knowing where a name starts is easy in Icelandic as they all start with a capital letter, but where do they end? Location names are usually only one word and names, middle names and last names for people all start with capital letters. But the company names are trickier, even though they usually start with a capital letter the following words can be any combination of numbers or words starting with capital, lower letter or even a number: ”Hafið bláa hafið ehf”,” Himinn og Haf auglýsingastofa ehf”, “66° Norður”. The system also has to be able to recognize foreign names and they do not follow these rules. Person names can include prepositions written in lower case such as “van”, “von” and “de”. Location names are often a combination of more than one word and can also include words in lower cases, for example, “Rio de Janiero”.

A simple but yet an effective way to recognize names is to use pattern matching. The Icelandic endings “-son” and “-dóttir”, (meaning son of, daughter of) will, unless associated with abbreviations such as hf, ehf or EA-11 (where the two first would point to a company name and the last one to a ship name), be enough to identify for certain a name of a person. The abbreviations mentioned above, “hf” and “ehf”, will always be part of a company name. Locations also have their typical endings referring to features of the landscape such as “-vatn”, “-fell” or “-fjall” (lake, hill, mountain) but unfortunately they are not as distinguishing as companies are often named after locations, “Arnarfell ehf”, “Básfell ehf”, “Laugafell ehf”, to name but a few.

Textual context can be of great help when categorizing names. Professional titles, for example, can stand in front of a person’s name if it has an article suffix and behind if no article is attached, “hagfræðingurinn Jón Jónsson” (economist-the Jón Jónsson), “Jón Jónsson hagfræðingur”(Jón Jónsson economist), but they can also be used to identify companies because if a title with no article stands in front of a named entity, then that entity can only be a company. “aðalhagfræðingur Seðlabankans, Jón Jónsson…” (main economist of the National bank, Jón Jónsson) . Words describing family relations such as “faðir, móðir, bróðir and systir” (father, mother, brother and sister) will very likely be standing close to a person’s name.

Gazette lists are often used with NER systems. They are a list of names which have been categorized and then used to look up entries. Such gazette lists of course vary in both size and content between systems and how they are utilized. Are they, for example, used as last resort or is the system totally dependent on them? Researchers have shown that their content is far more important than their size and a small gazette list with well known entries is far more helpful than a large list listing relatively unknown names which seldom appear in text (Mikheev, Grover, Moens 1999). If a name of a relatively unknown place will appear in a text it’s probably safe to assume that there will follow some explanation whereas the author of a text will not find any reasons to explain anything if he assumes that a name is generally well known to his readers. Another problem with relying too much on gazette list is the issue of “Form over Function” or “Function over Form”.

If, for example, Iceland plays a football match against Denmark it is not the locations Iceland and Denmark that are competing, but rather the organizations or companies (national teams) that play the game. When using “Form over Function” the same tag is used for all occurrences of a named entity, whereas the same named entity can be tagged differently if using “Function over Form”. Iceland could, for example, be tagged both as LOCATION and COMPANY depending on in what context it appears.

- Reykjavík is the capital of Iceland (Location)
- Iceland plays Denmark next Saturday (Company)

Many NER systems have been developed for specific languages, such as SweNam (Dalians, Åström 2001) for Swedish, DanGram (Bick 2004) for Danish and LTG (Mikheev, Grover, Moens 1998) for English, which was the highest scoring system in the Message Understanding Conference (MUC-7), held in 1997 (Marsh, Perzanowski 1998). In contrast, many other NER systems are language independent, for example, CMP02 (Carreras, Márques, Padró 2002) and Flo02 (Florian 2002). But just how accurate are these systems? In MUC-7 the F–score for English text ranged from 69,67% to 93,39%, calculated in the following way:

- **F-score:** 2 * Precision * Recall / (Recall + Precision)
- **Precision:** percentage of named entities found by the algorithm that are correct
- **Recall:** percentage of named entities defined in the corpus that were found by the program

Two human annotators scored 97,60% and 96,95% on the same occasion making the highest system score even more impressive (Marsh, Perzanowski 1998).

In the Conference on Computational Natural Language Learning 2002 (CoNLL-2002) the highest scoring system was CMP02 with an F-score of 81,39% for Spanish and 77,05% for Dutch, while Flo02 scored 79,05% for Spanish and 74,99% for Dutch (Universiteit Antwerpen n.d).

This system will categorize names into four groups; Company, Event, Location and Person. The Company group includes; companies, government organizations, political parties, sport clubs, etc. In the Event category will be names of festivals, shows, award ceremonies and competitions. The Location group is for names of places, countries, cities, etc. Names of persons, real or fictional, will go into the Person group. Not all names will be categorized. Titles, for example, movies, books, songs and albums do not fit into any of the above categories and will not be marked. Nationalities, names of items and brand names will also go unmarked.

2. Related Work

NER systems can roughly be categorized into either linguistic based or machine learning based systems. The linguistic based systems rely on handcrafted rules and are language dependent and the machine learning ones use corpora to learn from, either by supervised or unsupervised learning. The latter ones can be language independent which of course is a great advantage. Finally, there are systems that use a combination of both rules and statistics.

There exists to my best knowledge one other NER system for Icelandic, “Íslenskur textaskimi” who marks proper names like person, company and location in text but also looks for keywords (Rannís 2006). I did not find a lot of information about this system but from what I could gather it uses gazette lists and regular expressions to find and mark words and Named entities (Já 2006)

To try to make some comparison of the result I will use results presented in the paper “*Named Entity Recognition for the Mainland Scandinavian Languages*” (Johannessen, Hagen, Haaland, Jónsdottir, Kokkinakis, Meurer, Bick, Haltrup 2005). The paper describes a comparison of six NER systems; Norwegian CG (a rule-based system, based on constraint grammar), Norwegian ME (maximum entropy), Norwegian MBL (memory-base learning), Swedish FS (context-sensitive finite-state grammars), Danish FS and Danish CG.

2.1. Linguistic NER systems

As the name suggests, linguistic NER systems are based on linguistic rules. These rules can be as simple as case matching, recognizing suffixes or known combinations of words such as “hlutabréf í Xxx … “ (shares in Xxx) where the word “hlutabréf” can only be connected to a company. But they can also be more complicated taking into account morphological rules like that a named entity with an article in a genitive case following a professional title is a company name. It is, I think, safe to assume that grammatical knowledge of the languages is a major factor in how successful any linguistic NER system will be.

An example of a linguistic based system is the LTG system (although it also uses some machine learning methods) (Mikheev, Grover, Moens 1998). Its input is a tokenized and tagged text on which the LTG system applies a series of rules:

1. **Sure-fire**, which are context oriented and fire only when a possible candidate expression is surrounded by a suggestive context (Xxxx+ is a? JJ* PROF; “Yuri Gromov is a former director”; PERS). Meaning that if a named entity, (Xxxx) is followed by “is”, an optional “a”, zero or more adjectives and a professional title, as in “Yuri Gromov is a former director”. Then that entity is marked as person.
2. **Partial match 1**, takes entities already found in the document and generates all possible partial orders of the composing words preserving their order, it then uses a pre-trained maximum entropy model, based on contextual information.
3. **Relaxed rules**, again are context oriented, but this time more relaxed and extensively use the information from already existing markup and lexicons.
4. **Partial match 2**, similar to Partial match 1.
5. **Title assignment**, Newspaper titles are commonly written in capital letters, they are matched or partially matched with entities already found, and checked against a maximum-entropy model trained on document titles.

The LTG system was the highest scoring system in MUC 7, with an F-score of 93,39% (Marsh, Perzanowski 1998)

SweNam (Dalians, Åström 2001) is a Swedish NER system that uses mix of rules, lexicons (gazette list) and training strategies. The rules are made up of case matching and can be either for learning and matching, or matching only. An example of these rules are;

- Locations
    - The last word of a name contains a place ending e.g. -vägen -väg, -gatan, -gata, -parken, -park, etc, (Matching and Learning)

- Companies and Organations
    - More than one capital letter in a row e.g. AU-systems (Matching)
    - A company ending of the last word e.g. fabriken (factory) as in Framtidsfabriken (Matching and Learning)
- Person
    - Middle words in lower case e.g. Hans van der Vriees (Matching)
    - title(s) e.g. Mr Erik Åström, Vice VD Erik Åström (Matching)
- Time
    - formal date in many forms e.g. den 10:e januari 2001, 10:e januari, 10 januari, 2001 (Matching)

SweNam also uses lists that contains suffixes of locations and organization and prefixes, titles for persons.

- Locations
    - -gatan, -området, -torget, etc
- Organizations
    - -firma, -byrå, -företaget, etc
- Title lists
    - Mr, Mrs, Miss, Herr, Fru, etc

To evaluate SweNam 100 manually tagged texts were used, containing 1800 names and time, no mention of how many words (Dalians, Åström 2001).

Recall	38%
Precision	75%
F-Score	51%

2.2. Machine learning NER systems

The two main techniques for machine learning are supervised learning and unsupervised learning.

2.2.1. Supervised learning

In recent years, work on NER systems has been shifting towards supervised learning. In MUC-7 (1997), for example, five out of eight systems were rule based but in CONLL-2003 all sixteen systems where based on supervised learning (Nadeau 2007). It requires a manually annotated training corpus, the bigger the better, but its domain also plays a major role. Having a corpus containing news articles on finance will not necessarily prepare a NER system to recognize named entities from a sports article, no matter how big it is.

Typically these systems read the corpus, identify the name entity and try to establish disambiguation rules derived from the corpus.

One supervised learning method is Decision Tree (Sekine 1998) where a training corpus is read by the system which then builds a decision tree. It marks the tokens as opening, continuing, closing or none. If there are 8 name types then there are 33 possible outputs where output would be preceded with, for example, org, loc or date.

Output	**beginning of token**	**ending of token**	**token is**
OP-CL	opening	closing	NE itself
OP-CN	opening	continuing	starting NE
CN-CN	continuing	continuing	middle of NE
CN-CL	continuing	closing	ending NE
None	none	none	none

Table 1

Each leaf holds all possible tags for that word and their probabilities. Then when the text that is being marked is read, each word is looked up in the decision tree and its nametag decided based on the current, previous and following word. When all tokens in a sentence have been tagged the most probable consistent path through that sentence is chosen, meaning that there cannot be a combination of org-OP-CN, date-OP-CL where as loc-OP-CN, loc-CN-CL is very likely.

Other methods include Hidden Markov Models (Bikel, Miller, Schwartz, Weischedel 1997), Maximum Entropy Models (Borthwick, Sterling, Agichtein, Grishman 1998), Support Vector Machine (Asahara, Matsumotol 2003) and Conditional Random Fields (MCCallum, Li 2003).

2.2.2. Unsupervised learning

Not all machine learning methods require a large annulated corpus. In Named Entity Discovery Using Comparable News Article (Shinyama, Sekine 2004) the authors describe a method based on comparison. The idea is that newspaper articles from various newspapers on the same date will more or less be reporting on the same stories. As it is difficult to paraphrase names they should appear in all papers, the frequency of the names and distribution should be similar.

3. System description

As no annotated Icelandic corpus with marked named entities exists the choice is between creating one and then build a machine learning system or just build a linguistic system. I have chosen to go for a linguistic system, using a similar approach as the LTG system. The system reads the text several times, applying the strictest rules first and then more relaxed rules, the system will learn from some of the rules while others are for matching only.

The linguistic systems also have the advantage that it is possible to start with a small set of rules and then gradually build on it. As time is a limited resource, this feature makes the choice even easier.

The system will be built on two subsystems, the first one, called NameScanner, will use regular expressions to create lists of named entities based on endings such as “-son”, “-dóttir” and abbreviations like “hf”, “ehf”. It will also generate lists of words that can be of significance, such as professional titles, words that imply a location, a company or a person, etc.

The second subsystem called, NameFinder, will read these lists, and break up combinations of words if a name is made of more than a single word. If, for example, the name “Ingibjörg Sólrún Gísladóttir” appears in the name list, then entries for “Ingibjörg Sólrún”, “Ingibjörg”, “Sólrún”,“Gísladóttir” and “Sólrún Gísladóttir” will also be added. The NameFinder will also read the text itself, after it has been run through a Part-of-speech tagger (PoS). The tagger that will be used, IceTagger (Loftsson, H. 2008), tags amongst other proper nouns which are good candidates for a start of a named entity. IceTagger marks the proper nouns it finds as person name, location name and other. These markups are based on the data used when IceTagger was trained and no new names are learned. This markup will not be used by the NameFinder. The NameFinder will then use the name lists and rules based on, for example, the context in which these entities appear to try to categorize these entities. These rules will not be equal, some will give a 100% certainty while others will only give a vague indication.

The NameFinder is capable of carrying out some declination of nouns and, for example, knows that “a” umlauts under certain circumstances into “e”, “ö” and vice versa. It also knows the case endings listed in the following tables and is therefore capable of recognizing, for example, “Vörður”, “Vörð ”, “Verði” and “Varðar” or “Hildur”, “Hildi”, “Hildi” and “Hildar” as the same named entity.

Masculine, no article						
nominative	i	-	ur	-	-	n
accusative	a	-	-	-	-	-
dative	a	i	i	-	i	i
genitive	a	ar	ar	s	s	s

Table 2

Feminine, no article				
nominative	-	-	ur	a
accusative	-	u	i	u
dative	-	u	i	u
genitive	ar	ar	ar	u

Table 3

There is also the option to use predefined gazette lists and they can include both named entities followed by a tag and roles followed by a tag. For the named entities it is not necessary to list all declined forms of a noun as the NameFinder recognizes the ones listed above, but for roles it would be preferable to list all possible variants of the word. The form of the gazette lists is, where SEP stands for seperator:

- London SEP LOCATION
- Borgarfulltrúi SEP ROLE_PERSON
- Móðir SEP RELATION_PERSON

The difference between role and relation is that role can be mapped to another tag while a relation tag is always used as the name tag.

At runtime the user has the option to run the NameFinder system with or without a gazette list and in default or greedy mode. In a greedy mode, all unmarked named entities that follow the preposition “á” and “í” are marked as locations and names with the pattern “Xxxx Xxxx” are marked as persons. To run the NameFinder system from the command prompt

• Default	C:\ [dir] >NameFinder <input file> <output file>
• Default with gazette list	C:\ [dir] >NameFinder <input file> <output file> <gazette file>
• Greedy	C:\ [dir] >NameFinder <input file> <output file> -g
• Greedy with gazette list	C:\ [dir] >NameFinder <input file> <output file> <gazette file> -g

I don’t think I have any new ideas as to how to recognize named entities. My Role/Relation rules are similar to the sure-fire rules of the LTG system (Mikheev, Grover, Moens 1998) and I use endings as many systems before, for example, SweNam (Dalians, Åström 2001).

If, for example, the following text would be fed to the system “*Er Árni Þór Sigurðsson þingmaður Vinstri grænna í raun Soffía frænka?*“ *(Is Árni Þór Sigurðsson mp Vinstri grænna in fact Soffía aunt)* the output from the NameScanner would be:

- Er Árni Þór Sigurðsson SEP PERSON
- þingmaður SEP ROLE_PERSON
- frænka SEP RELATION_PERSON

Where the underlined endings are recognized by the NameScanner.

IceTagger:

Er sfg3en	*(verb, indicative, active, 3rd person, singular, present)*
Árni nken-m	*(noun, masculine, singular, nominative, no suffixed, person name)*
Þór nken-m	*(noun, masculine, singular, nominative, no suffixed, person name)*
Sigurðsson nken-m	*(noun, masculine, singular, nominative, no suffixed, person name)*
þingmaður nken	*(noun, masculine, singular, nominative)*
Vinstri lheevf	*(adjective, neuter, singular, genitive, weak, positive)*
grænna lheevm	*(adjective, neuter, singular, genitive, weak, comparative)*
í aþ	*(adverb and preposition, governs dative)*
raun nveþ	*(noun, feminine, singular, dative)*
Soffía nven-m	*(noun, feminine, singular, nominative, no suffixed, person name)*
frænka nven	*(noun, feminine, singular, nominative)*
? ?	

IceTagger categorizes the proper nouns it finds as person name, location name and other. These markups are based on the data used when IceTagger was trained and no new names are learned. Here IceTagger correctly marks Árni, Þór, Sigurðsson and Soffía as person names but outher names might be missed. This markup will not be used by the NameFinder.

And finally from the NameFinder:

Er [Árni Þór Sigurðsson PERSON] þingmaður [Vinstri grænna COMPANY] í raun [Soffía PERSON] frænka ?

Where Árni Þór Sigurðsson is recognized by the NameScanner, Vinstri grænna found by the personal role, þingmaður and Soffía by the personal relation frænka.

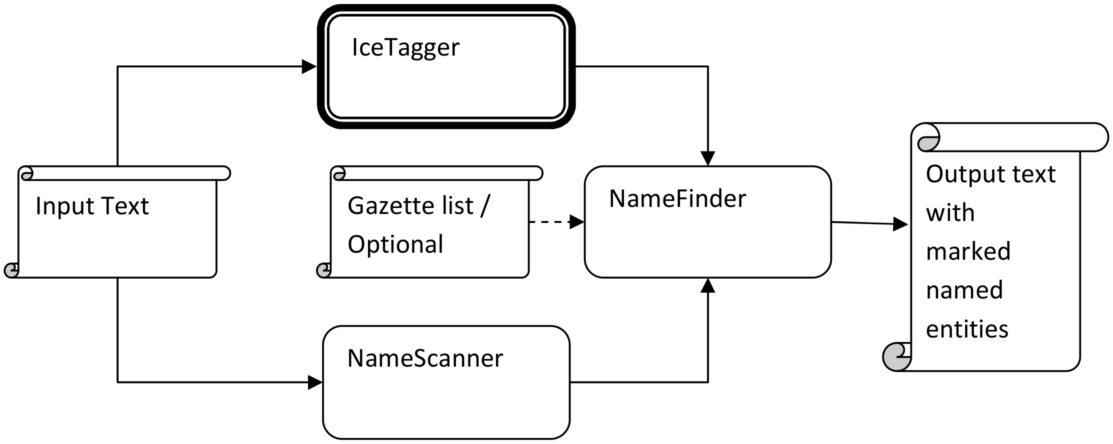


Figure 1

4. Implementation

The system is designed as two Java programs. The NameScanner is implemented in JFlex (Klein 2005). It uses regular expressions to mark words as names for persons, companies or locations based on pattern matching and endings, where, for example, "Xxxx xxx hf" would be marked as a company and Xxxfell as a location. It also marks words that can be helpful when recognizing named entities as person, company or location roles, for example, "hagfræðingur Person Role, faðir Person Relation, veitingahús Company Role and borg Location Role" (economist, father, restaurant and city). The NameScanner then prints out a list containing the words it found and their tag.

The NameFinder reads two separate files; the list constructed by NameScanner and the text after it has been fed through a Part-of-Speech tagger (IceTagger). The NameFinder constructs two tables, one for names and another for roles and relations, from the NameScanner file. For the names table, it also splits up combined names and adds sub combinations of that name into the table.

The NameFinder marks all proper nouns, foreign words starting with a capital letter and words in the middle of a sentence starting with a capital letter, as a possible start of a named entity. It then looks at words at the beginning of a sentence, if they have already been identified as a start of a name elsewhere in the text, then they are considered to be a start of a named entity. If they are not found in the same case, then the NameFinder checks if they exist in a different case.

To find the full name, NameFinder checks the words that follow and accepts the following combinations, if the start word is a proper noun or a foreign word.

- proper noun in the same case or foreign word
  - Ingibjörg Sólrún Gísladóttir
- genitive noun + genitive noun + conjunction + genitive noun
  - Bandalag starfsmanna ríkis og bæja
  - (Alliance of employees of state and towns)
- genitive adjective + genitive noun
  - Samband íslenskra sveitarfélaga
  - (Coalition of Icelandic municipalities)
- genitive noun
  - Bandalag háskólamanna
  - (Association of Academics)

- adjective
    - Hafið bláa
    - (Ocean-the blue)

If the first word in a name is an adjective

- Adjective
  - Vinstri grænir
  - (Left green)
- Noun
  - Hvíta húsið
  - (White house-the)

If that name is then identified as event, person, company or location name, then all the words that make up that name, get either a PERSON, COMPANY, EVENT or LOCATION name tag.

To finish partially found company names, the NameFinder looks for company endings like "hf, ehf, and ltd" and marks them and the preceding (max three) unmarked words to the nearest known company name.

The role/relations are the core of the NameFinder. They can either be preceding, following or a part of the name. If a role is following then its tag is simply used as the name tag. If a role is preceding then it depends on, for example, if it has an article ending, whether the role tag itself is used or it is mapped to an another tag. Names that are found by these rules are also added to the name table. The following list enumerates possible combinations of role/relation and names.

- Relation - pronoun - name
  - Lebedev keypti skuldum vafið blaðið í janúar og er hann meðeigandi þess ásamt (**syni** RELATION_PERSON) (**sínum** pronoun) [**Evgeny** PERSON].
  - (Leedev bought debt enveloped newspaper-the in January and is he partner its along **son his Evgeny.**)
- Relation - name
  - (**Pabbi** RELATION_PERSON) [**Britney** PERSON] heldur verndarhendi yfir dóttur sinni eftir að hún spilaði út fyrir um ári síðan.
  - (Father Britney holds protecting-hand over daughter his after she played out for a year ago.)
- Name - ,? - relation
  - Er Árni Þór Sigurðsson þingmaður Vinstri grænna í raun [**Soffía** PERSON] (**frænka** RELATION_PERSON)?

    - (Is Árni Þór Sigurðsson mp Vinstri grænna in fact Soffía aunt?)

- Role - name - , - name
    - (**hagfræðingur** ROLE_PERSON) (**Straums fjárfestingabanka** name), [**Raffaella Tenconi** PERSON].
    - (Economist ROLE_PERSON) (Straums financial-bank name) [Raffaella Tenconi PERSON].
- Role - noun - ,? - Name
    - Búningurinn sem (**konungur** ROLE_PERSON) (**poppsins** noun), [**Michael Jackson** PERSON], klæddist.
    - Outfit-the that (king ROLE) (pop noun) [Michael Jackson PERSON], wore.
- Role - ,? - name
    - Brady Dougan, (**forstjóri** ROLE_PERSON) [**Credit Suisse** COMPANY], segir hagnaðinn sýna að þær aðgerðir sem bankinn hafi gripið til séu að skila árangri.
    - Brady Dougan (manager ROLE_PERSON) [Credit Suisse COMPANY], says profit-the show that the actions that bank-the has ….
    - (**Fegurðardrottningin** ROLE_PERSON) [**Carrie Prejean** PERSON], sem varð í öðru sæti í keppninni Ungfrú Bandaríkin.
    - (Beuty-queen-the ROLE_PERSON) [Carrie Prejean PERSON] that came in second place in contest-the Miss USA.
- Name - ,? - role
    - [**Brady Dougan** PERSON], (**forstjóri** ROLE_PERSON) Credit Suisse, segir hagnaðinn sýna að þær aðgerðir sem bankinn hafi gripið til séu að skila árangri.
    - [Brady Dougan PERSON] (manager ROLE_PERSON) Credit Suisse, says profit-the show that the actions that bank-the has ….
- Name - ,? - adjective - role
    - Rússnesk stjórnvöld halda enn opnum þeim möguleika, að koma Iskander flugskeytum fyrir í [**Kalíningrad** LOCATION], (**rússnesku** adjective) (**landsvæði** ROLE_LOCATION) á milli Póllands og Litháens.

    - Russian officials hold still open the option, to place Iskander missiles in [Kaliningrad Location], (Russian adjective) (district ROLE_LOCATION) in between Poland and Lithuania.

- Role - part of name
    - Raffaella Tenconi, hagfræðingur hjá [**Straumi** (**fjárfestingabanka** ROLE_COMPANY) COMPANY] í London.
    - Raffaella Tenconi, economist with [Straum financial-bank ROLE_COMPANY) COMPANY] in London.

Abbreviations in all uppercase not separated by a dot are likely named entities. If they are found, already marked names are checked and if a name is found that contains the same initials, the abbreviation is considered to match that name and gets the same nametag. If no match is found they are simply marked as company names.

The NameFinder stores all given names, for persons, that it finds and then compares them to all already marked persons names, which stand alone as the second word in a sentence. If the second word is not a given name and the first word in the sentence is not known as a role or relation, then the second word is considered to be a family name and the preceding word is tagged as a person name.

Listings and pairs can give a clue if one or more of the named entities has already been tagged, “Stórfyrirtæki á borð við **Baug** og **FL Group** styrktu frambjóðendur...”(Big-companies like Baug and FL Group funded nominees). Here, for example, “Baug” and “FL Group” are very likely of the same kind. The same goes for this listing “… og má þar nefna fyrirtækin **Eff2 technologies**, **CLARA**, **Bjarmalund** , **Tunerific** og **Vinun**.” (and to name company-the Eff2 technologies , CLARA, Bjarmalund , Tunerific og Vinun.) A list of names is considered to be a listing if a full name is followed by a comma, a full name, conjunction or comma and then full name and an end token. A pair is simply a full name, conjunction and a full name. Then when printing out the marked text it is simply a question of adding an opening mark in front of the first word that has a nametag and then that nametag and a closing mark for the last word for each name.

5. Evaluation

The system will be measured with and without gazette list and in greedy and non-greedy mode. The gazette list includes only location names; the continents, countries, the states of USA, capital cities and the biggest cities of Europe and USA. All in all there are 523 entries in the list. To create the list I used the Icelandic version of the Wikipedia website, and searched for the countries, states and cities.

To measure the accuracy of the NameFinder system, news articles from mbl.is (2.5.2008) were used. They were split up into five categories; domestic news, foreign news, financial news, lifestyle and sports. Each category was tested separately and finally they were all combined in a final test. In total the test data contained 8212 words and 496 names. The result are presented as:

- $\text{Recall} = \frac{\text{total number of correctly found names}}{\text{total number of names}}$
- $\text{Precision} = \frac{\text{total number of correctly found names}}{\text{total number of names found}}$
- $\text{F} - \text{Score} = \frac{2 * \text{Recall} * \text{Precission}}{\text{Recall} + \text{Precission}}$

In advance, I would assume that the score will be highest in the domestic news category but lowest for the sports and lifestyle articles. The reason being that the system relies on persons to be introduced with a full name and a title. Although that seems to be the norm in news articles this introduction is often lacking in sports and lifestyle articles. Another problem is that sport clubs are often named after locations, for example, "Grindavík, Keflavík and Njarðvík" are all teams that play in the Icelandic basketball top league, named after their hometowns. In the lifestyle articles the content often includes titles of movies, songs, etc. that may include a person's name but should not be tagged at all.

In total there are 496 names in the test data, divided into roughly 1/3 Company names, 1/3 Locations and 1/3 Person names, but the exact division varies greatly between categories.

To try to make some comparison to other similar systems I will compare the result of the NameFinder system to results presented in the paper "*Named Entity Recognition for the Mainland Scandinavian Languages*" where six systems for Danish, Norwegian and Swedish were discussed (Johannessen, Hagen, Haaland, Jónsdottir, Kokkinakis, Meurer, Bick, Haltrup 2005). News articles in each language were used to evaluate the systems. The test data contained 1800 words and the number of named entities varied from language to language. Five of the systems mark names as:

• PRS	Person
• ORG	Organization
• LOC	Location
• WRK	Work of art
• EVT	Event
• SUM	The total of the above

The Danish FS marks names as:

- PRS	Person
- ORG	Organization
- LOC	Location
- SUM	The total of the above

I will then move on to a more in depth analysis of each category; domestic news, foreign news, financial news, lifestyle and sports for the NameFinder system. There are no separate results for the sub categories in the paper “*Named Entity Recognition for the Mainland Scandinavian Languages*”, so there will be no comparisons for the categories separately. All six of the Scandinavian systems used gazette lists, the number of names varied from 13.120 – 68,390. The Norwegian systems used a gazette list containing 13.120 names, the Danish FS 28.700 names, The Danish CG 44.200 and finally the Swedish FS 68.390 names (Johannessen, Hagen, Haaland, Jónsdottir, Kokkinakis, Meurer, Bick, Haltrup 2005). The gazette list used for the NameFinder system contained 523 names which seems rather modest compared to the other systems. The paper “*Named Entity Recognition for the Mainland Scandinavian Languages*” also presented results, without gazette lists, for recall for The Swedish FS and the Norwegian CG but no results for precisions or F-Score.

The Icleandic NameFinder system, in default mode						
	Correctly found	Total found	Total no.	Recall	Precision	F-Score
PERSON	106	114	153	69,28%	92,98%	79,40%
COMPANY	98	144	174	56,32%	68,06%	61,64%
LOCATION	61	67	163	37,42%	91,04%	53,04%
EVENT	2	3	6	33,33%	66,67%	44,44%
TOTAL	267	328	496	53,83%	81,40%	64,81%

Table 5

The Icleandic NameFinder system, in default mode with gazette list						
	Correctly found	Total found	Total no.	Recall	Precision	F-Score
PERSON	106	114	153	69,28%	92,98%	79,40%
COMPANY	93	110	174	53,45%	84,55%	65,49%
LOCATION	131	167	163	80,37%	78,44%	79,39%
EVENT	2	3	6	33,33%	66,67%	44,44%
TOTAL	332	394	496	66,94%	84,26%	74,61%

Table 6

The Icleandic NameFinder system, in greedy mode						
	Correctly found	Total found	Total no.	Recall	Precision	F-Score
PERSON	134	182	153	87,58%	73,63%	80,00%
COMPANY	102	149	174	58,62%	68,46%	63,16%
LOCATION	100	116	163	61,35%	86,21%	71,68%
EVENT	2	3	6	33,33%	66,67%	44,44%
TOTAL	338	450	496	68,15%	75,11%	71,46%

Table 7

The Icleandic NameFinder system, in greedy mode with gazette list						
	Correctly found	Total found	Total no.	Recall	Precision	F-Score
PERSON	133	156	153	86,93%	85,26%	86,08%
COMPANY	97	116	174	55,75%	83,62%	66,90%
LOCATION	150	193	163	92,02%	77,72%	84,27%
EVENT	2	3	6	33,33%	66,67%	44,44%
TOTAL	382	468	496	77,02%	81,62%	79,25%

Table 8

The highest F-Score when all categories are combined is 79,25%, obtained in greedy mode with a gazette list (see tables 5–8). This would rank the NameFinder system third compared to the Scandinavian systems (see tables 9–14). As there is very limited information on results without a gazette lists one can only assume that the NameFinder system would rank between the Norwegian CG, recall of 83% (see table 16) and the Swedish FS, recall of 53,29% (see table 10), with a recall of 68,15% (see table 7). The results for precision are missing so there is no way to calculate the F-Score for the Norwegian CG and Swedish FS but The NameFinder system had an F-Score of 71,46% (see table 7).

The NameFinder's weakest category of names is company names where it scores from 61,64% to 66,90% (see tables 5-8) and events with an F-Score of 44,44%. The events are relatively few or only 1,21% of the named entities so they do not have a great impact on the overall result. In default mode the F-Score is only 53,04% for locations (see table 5) but in other modes the F-Score is from 71,68% to 84,27% (see tables 6-8). The main reason for errors is that company names are wrongly marked as locations (mostly sport clubs, named after cities). This not only lowers the recall for companies but also lowers precisions for locations resulting in a lower F-Score for both categories. This can be seen if we compare precision for locations with and without gazette list were it drops from 91,04% to 78,44% (see tables 5-6) in default mode, when gazette list is added and from 86,21% to 77,72% (see tables 7-8) in greedy mode.

The greedy mode lives up to its name and finds considerably more names than the default mode or 450 versus 328 without gazette list (see tables 5 and 7) and 468 versus 394 with gazette list (see tables 6 and 8). This results in higher recall 53,83% to 68,15% without gazette list (see tables 5 and 7) and 66,94% to 77,02% with gazette list (see tables 6 and 8), but also in lower precision 81,40% to 75,11% without gazette list (see tables 5 and 7) and 84,26% to 81,62% with gazette list (see tables 6 and 8). As a result the F-Score rises from 64,81% to 71,46% without gazette list (see tables 5 and 7) and from 74,61% to 79,25% with gazette list (see tables 6 and 8).

The Swedish FS system, with gazetteers

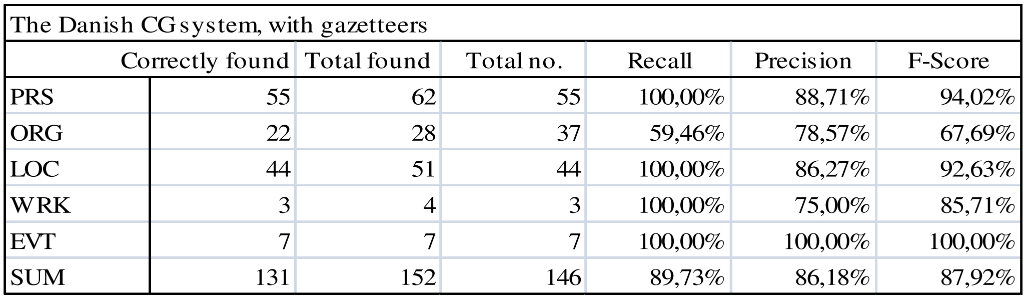
	Correctly found	Total found	Total no.	Recall	Precision	F-Score
PRS	68	75	71	95,77%	90,67%	93,15%
ORG	24	27	30	80,00%	88,89%	84,21%
LOC	35	36	38	92,11%	97,22%	94,59%
WRK	6	6	8	75,00%	100,00%	85,71%
EVT	5	5	5	100,00%	100,00%	100,00%
SUM	138	149	152	90,79%	92,62%	91,69%

Table 9

The Swedish FS system, without gazetteers

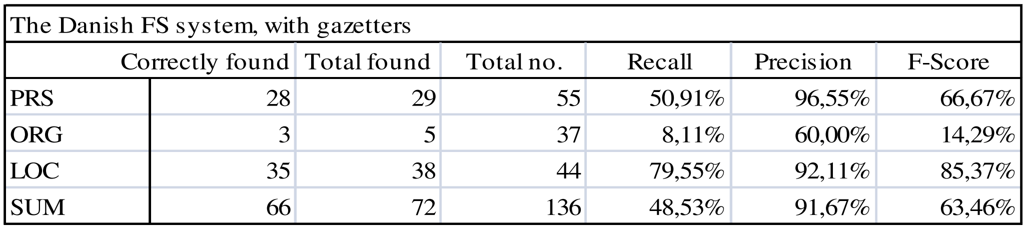
	Correctly found	Total found	Total no.	Recall	Precision	F-Score
PRS	42		71	59,15%		
ORG	15		30	50,00%		
LOC	13		38	34,21%		
WRK	7		8	87,50%		
EVT	4		5	80,00%		
SUM	81		152	53,29%		

Table 10



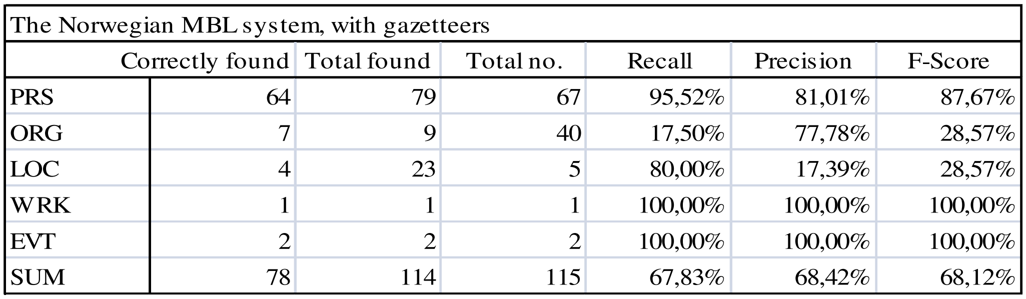
The Danish CG system, with gazetteers						
	Correctly found	Total found	Total no.	Recall	Precision	F-Score
PRS	55	62	55	100,00%	88,71%	94,02%
ORG	22	28	37	59,46%	78,57%	67,69%
LOC	44	51	44	100,00%	86,27%	92,63%
WRK	3	4	3	100,00%	75,00%	85,71%
EVT	7	7	7	100,00%	100,00%	100,00%
SUM	131	152	146	89,73%	86,18%	87,92%

Table 11



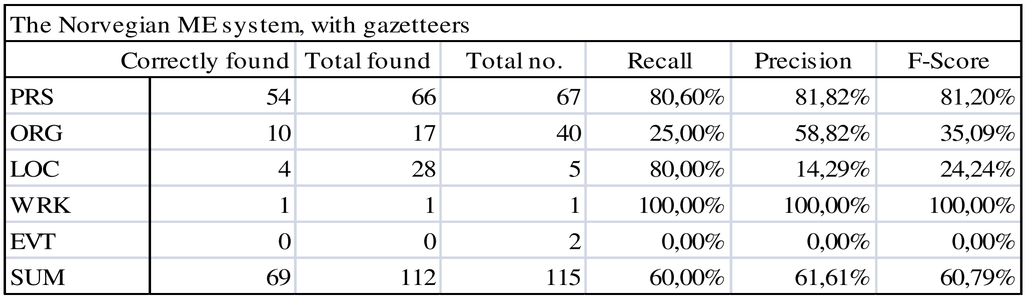
The Danish FS system, with gazetters						
	Correctly found	Total found	Total no.	Recall	Precision	F-Score
PRS	28	29	55	50,91%	96,55%	66,67%
ORG	3	5	37	8,11%	60,00%	14,29%
LOC	35	38	44	79,55%	92,11%	85,37%
SUM	66	72	136	48,53%	91,67%	63,46%

Table 12



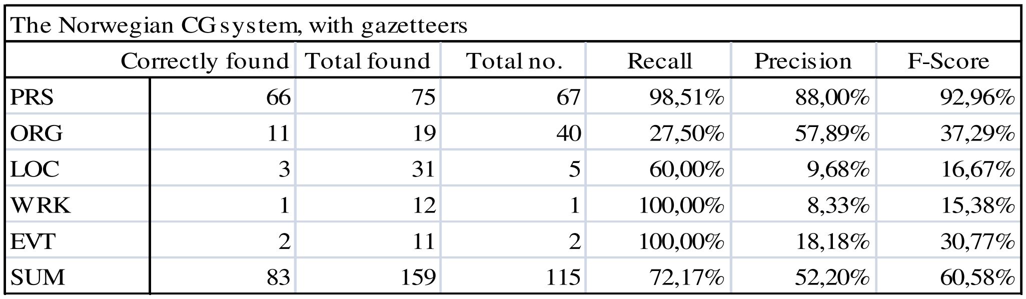
The Norwegian MBL system, with gazetteers						
	Correctly found	Total found	Total no.	Recall	Precision	F-Score
PRS	64	79	67	95,52%	81,01%	87,67%
ORG	7	9	40	17,50%	77,78%	28,57%
LOC	4	23	5	80,00%	17,39%	28,57%
WRK	1	1	1	100,00%	100,00%	100,00%
EVT	2	2	2	100,00%	100,00%	100,00%
SUM	78	114	115	67,83%	68,42%	68,12%

Table 13



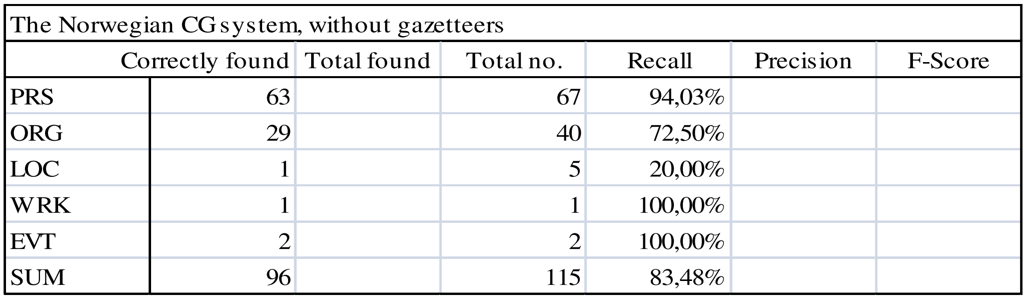
The Norvegian ME system, with gazetteers						
	Correctly found	Total found	Total no.	Recall	Precision	F-Score
PRS	54	66	67	80,60%	81,82%	81,20%
ORG	10	17	40	25,00%	58,82%	35,09%
LOC	4	28	5	80,00%	14,29%	24,24%
WRK	1	1	1	100,00%	100,00%	100,00%
EVT	0	0	2	0,00%	0,00%	0,00%
SUM	69	112	115	60,00%	61,61%	60,79%

Table 14



The Norwegian CG system, with gazetteers						
	Correctly found	Total found	Total no.	Recall	Precision	F-Score
PRS	66	75	67	98,51%	88,00%	92,96%
ORG	11	19	40	27,50%	57,89%	37,29%
LOC	3	31	5	60,00%	9,68%	16,67%
WRK	1	12	1	100,00%	8,33%	15,38%
EVT	2	11	2	100,00%	18,18%	30,77%
SUM	83	159	115	72,17%	52,20%	60,58%

Table 15



The Norwegian CGsystem, without gazetteers						
	Correctly found	Total found	Total no.	Recall	Precision	F-Score
PRS	63		67	94,03%		
ORG	29		40	72,50%		
LOC	1		5	20,00%		
WRK	1		1	100,00%		
EVT	2		2	100,00%		
SUM	96		115	83,48%		

Table 16

The results of the NameFinder system are presented in the following tables seperately for the five news categories: domestic news, foreign news, financial news, lifestyle and sports. The results for each category are discussed and the errors analyzed.

Domestic news		Company	Event	Location	Person	Total
Default	Recall	87,10%	0,00%	83,33%	82,35%	84,21%
	Precision	96,43%	0,00%	96,15%	90,32%	94,12%
	F-Score	**91,53%**	**0,00%**	**89,29%**	**86,15%**	**88,89%**
Default with gazette list	Recall	87,10%	0,00%	86,67%	82,35%	85,26%
	Precision	96,43%	0,00%	96,30%	90,32%	94,19%
	F-Score	**91,53%**	**0,00%**	**91,23%**	**86,15%**	**89,50%**
Greedy	Recall	87,10%	0,00%	90,00%	88,24%	88,42%
	Precision	96,43%	0,00%	93,10%	85,71%	91,30%
	F-Score	**91,53%**	**0,00%**	**91,53%**	**86,96%**	**89,84%**
Greedy with gazette list	Recall	87,10%	0,00%	93,33%	88,24%	89,47%
	Precision	96,43%	0,00%	93,33%	85,71%	91,40%
	F-Score	**91,53%**	**0,00%**	**93,33%**	**86,96%**	**90,43%**

Table 17

Domestic news	
Word count	2923
Total number of names	95
Company	31
Event	0
Location	30
Person	34

Table 18

In the Domestic news category the NameFinder system scores reasonably well (see table 17). It benefits less than 1% from the use of gazette lists and highest accuracy is obtained in greedy mode with gazette list. The named entities are pretty evenly divided between Company, Location and Person and there is no Event (see table 18).

The main cause of errors in this category is that names are only partly found. There are, for example, two references to “Ásta Ragnheiður Jóhannesdóttir” and one to “Ásta Ragnheiði” and in all cases at the start of the sentence and tagged as a regular noun by IceTagger. In all three occurrences the system misses “Ásta” but correctly tags “Ragnheiði Jóhannesdóttur” and “Ragnheiði” as a person. If the name “Ásta” would have appeared once in the middle of a sentence, that would have been enough to correctly complete all three occurrences, as “Ásta” would then have been a known start of a name. The band “The Virgin Tongues” appears once. “The” is tagged as a company but “Virgin Tongues” either skipped or tagged as a person. The reason for that is that there is a requirement that if words are tagged as proper nouns, then they have to be in the same case to be considered as part of the same name. This rule may be too strict.

Foreign news		Company	Event	Location	Person	Total
Default	Recall	81,48%	0,00%	25,23%	82,76%	44,79%
	Precision	41,51%	0,00%	93,10%	96,00%	68,22%
	F-Score	**55,00%**	**0,00%**	**39,71%**	**88,89%**	**54,07%**
Default with gazette list	Recall	81,48%	0,00%	83,18%	82,76%	82,82%
	Precision	81,48%	0,00%	94,68%	96,00%	92,47%
	F-Score	**81,48%**	**0,00%**	**88,56%**	**88,89%**	**87,38%**
Greedy	Recall	81,48%	0,00%	49,53%	86,21%	61,35%
	Precision	41,51%	0,00%	94,64%	62,50%	67,11%
	F-Score	**55,00%**	**0,00%**	**65,03%**	**72,46%**	**64,10%**
Greedy with gazette list	Recall	81,48%	0,00%	93,94%	86,21%	90,18%
	Precision	81,48%	0,00%	95,24%	92,59%	92,45%
	F-Score	**81,48%**	**0,00%**	**94,34%**	**89,29%**	**91,30%**

Table 19

Foreign news	
Word count	2470
Total number of names	163
Company	27
Event	0
Location	107
Person	29

Table 20

The Foreign news category benefits the most from use of the gazette list (see table 19), not surprisingly as the gazette list includes names of countries and major foreign cities which appear frequently in foreign news. When the gazette list is introduced, the recall for locations rises from 25,23% to 83,18% and the F-Score from 39,71% to 88,56% (see table 19). Location names make up approximately 2/3 of the named entities in this category (see table 20), all of them foreign and most of them relatively well known. Therefore there are not many clues in the context as to what kind of a place they are (as would have been expected had a small city in Holland or the capital of Honduras been mentioned). "Mexíkó", for example, appears 17 times in this text, making up 16% of the Location category. When greedy mode is applied without the gazette list the recall for locations improves from 25,23% to 49,53% without any drop in precision, 93,10% and 94,64% (see table 19) so the rule that names following the prepositions "á" and "í" should be marked as Location is quite successful.

It is also noticeable that the accuracy for company names improves greatly when the gazette list is introduced. The F-score rises from 55,00% to 81,58% (see table 19). The reason for that is that location names are wrongly tagged as companies without the gazette list and the precision suffers severely, 41,51% for companies without the gazette list, versus 81,48% with the gazette list (see table 19). Again the highest F-score is obtained in greedy mode with gazette list.

Business news		Company	Event	Location	Person	Total
Default	Recall	66,67%	0,00%	71,43%	100,00%	71,64%
	Precision	91,89%	0,00%	83,33%	100,00%	92,31%
	F-Score	**77,27%**	**0,00%**	**76,92%**	**100,00%**	**80,67%**
Default with gazette list	Recall	60,78%	0,00%	85,71%	100,00%	68,66%
	Precision	91,18%	0,00%	54,55%	100,00%	85,19%
	F-Score	**72,94%**	**0,00%**	**66,67%**	**100,00%**	**76,03%**
Greedy	Recall	68,63%	0,00%	85,71%	100,00%	74,63%
	Precision	94,59%	0,00%	50,00%	75,00%	81,97%
	F-Score	**79,55%**	**0,00%**	**63,16%**	**85,71%**	**78,12%**
Greedy with gazette list	Recall	62,75%	0,00%	85,71%	100,00%	70,15%
	Precision	91,43%	0,00%	37,50%	75,00%	74,60%
	F-Score	**74,42%**	**0,00%**	**52,17%**	**85,71%**	**72,31%**

Table 21

Business news	
Word count	955
Total number of names	67
Company	51
Event	0
Location	7
Person	9

Table 22

In the Business news, company names are the vast majority, or 76,11% (see table 22). Here the gazette list actually lowers the score, as names like The "Washington Post" are wrongly associated with location and the F-score drops from 80,67% to 76,03% (see table 21). This is the only category where the default mode has the highest F-score, 80,67% (see table 21). The precision is quite good without the gazette list but out of 51 company names 17 are missed. This explains the relatively low recall of 66,67% for company names without gazette list (see table 21). Six of these missed company names belong to the same company, "Promens" which would, if caught, have greatly improved the recall. With gazette list the recall for company names drops to 60,78% (see table 21) as some company names are wrongly tagged as location.

The NameFinder system finds and correctly tags all nine person names in the default mode. Eight of these names are Icleandic and one is foreign, "Mark LaNeve".

Lifestyle news		Company	Event	Location	Person	Total
Default	Recall	18,18%	0,00%	0,00%	41,38%	31,11%
	Precision	50,00%	0,00%	0,00%	100,00%	82,35%
	F-Score	**26,67%**	**0,00%**	**0,00%**	**58,54%**	**45,16%**
Default with gazette list	Recall	18,18%	0,00%	75,00%	41,38%	37,78%
	Precision	50,00%	0,00%	50,00%	100,00%	73,91%
	F-Score	**26,67%**	**0,00%**	**60,00%**	**58,54%**	**50,00%**
Greedy	Recall	45,45%	0,00%	50,00%	79,31%	66,67%
	Precision	62,50%	0,00%	100,00%	74,19%	71,43%
	F-Score	**52,63%**	**0,00%**	**66,67%**	**76,67%**	**68,97%**
Greedy with gazette list	Recall	45,45%	0,00%	75,00%	79,31%	68,89%
	Precision	71,43%	0,00%	42,86%	79,31%	70,45%
	F-Score	**55,56%**	**0,00%**	**54,55%**	**79,31%**	**69,66%**

Table 23

Lifestyle news	
Word count	674
Total number of names	45
Company	11
Event	1
Location	4
Person	29

Table 24

In Lifestyle articles people are the main subject and person names make up for 64,44% of the named entities (see table 24). Therefore, the benefit of the gazette list is minor. Again the greedy mode scores higher than the default mode (see table 23). The system correctly tags “Heidi Klum” as a person in default mode but fails to learn that “Klum” is a person name and therefore misses the next six occurrences of “Klum” ,which count for 20% of the person names (see table 24), because the words “Heidi” and “Klum” are tagged in different cases. Again, the requirement that proper nouns have to be in the same case may be too strict. In greedy mode, “Klum” is learnt as a person name and properly tagged and the recall for person names rises from 41,38% to 79,31% (see table 23). The precision for person names drops from 100% to 74,19% (see table 23) when the greedy mode is applied as not all the names added are correct, but the F-score is still increased considerably and rises from 58,54% to 79,31% (see table 23). The F-score for company names is also greatly improved in greedy mode and rises from 26,67% to 55,56% (see table 23). The precision improves from 62,50% with the introduction of the gazette list as some locations were marked as companies without the gazette list.

Sport news		Company	Event	Location	Person	Total
Default	Recall	22,64%	40,00%	35,71%	64,00%	41,80%
	Precision	75,00%	100,00%	83,33%	91,43%	86,44%
	F-Score	**34,78%**	**57,14%**	**50,00%**	**75,29%**	**56,35%**
Default with gazette list	Recall	18,87%	40,00%	50,00%	64,00%	41,80%
	Precision	71,43%	100,00%	25,93%	91,43%	65,38%
	F-Score	**29,85%**	**57,14%**	**34,15%**	**75,29%**	**51,00%**
Greedy	Recall	22,64%	40,00%	92,86%	92,00%	59,84%
	Precision	66,67%	100,00%	72,22%	74,19%	73,00%
	F-Score	**33,80%**	**57,14%**	**81,25%**	**82,14%**	**65,77%**
Greedy with gazette list	Recall	18,87%	40,00%	92,86%	90,00%	57,38%
	Precision	62,50%	100,00%	37,14%	86,54%	66,67%
	F-Score	**28,99%**	**57,14%**	**53,06%**	**88,24%**	**61,67%**

Table 25

Sport news	
Word count	1190
Total number of names	122
Company	53
Event	5
Location	14
Person	50

Table 26

The names in the Sport news are pretty evenly split between companies and persons (see table 26). The F-scores for persons are quite decent in both modes, with and without a gazette list, but for companies the F-score is between ca 29% – 35% (see table 25) as the NameFinder misses 40 sport clubs names in default mode and then with the introduction of the gazette list tags previously found company names like “Manchester United” as location. The context in sport news is not particularly helpful because the language use tends to differ from the vocabulary and syntax used in domestic and foreign news. The NameFinder system is not able to use sentences like “*Miami Heat tók Atlanta Hawks í bakaríið í nótt og sigraði stórt*” to make any assumption as to whether “Miama Heat” and “Atlanta Hawks” are companies or not. When the gazette list is introduced both these entities are associated with locations.

Total		Company	Event	Location	Person	Total
Default	Recall	56,32%	33,33%	37,42%	69,28%	53,83%
	Precision	68,06%	66,67%	91,04%	92,98%	81,40%
	F-Score	**61,64%**	**44,44%**	**53,04%**	**79,40%**	**64,81%**
Default with gazette list	Recall	53,45%	33,33%	80,37%	69,28%	66,94%
	Precision	84,55%	66,67%	78,44%	92,98%	84,26%
	F-Score	**65,49%**	**44,44%**	**79,39%**	**79,40%**	**74,61%**
Greedy	Recall	58,62%	33,33%	61,35%	87,58%	68,15%
	Precision	68,46%	66,67%	86,21%	73,63%	75,11%
	F-Score	**63,16%**	**44,44%**	**71,68%**	**80,00%**	**71,46%**
Greedy with gazette list	Recall	55,75%	33,33%	92,02%	86,93%	77,02%
	Precision	83,62%	66,67%	77,72%	85,26%	81,62%
	F-Score	**66,90%**	**44,44%**	**84,27%**	**86,08%**	**79,25%**

Table 27

Total	
Word count	8212
Total number of names	496
Company	174
Event	6
Location	163
Person	153

Table 28

6. Conclusion and future work

When designing and implementing the system the main focus was on domestic and foreign news. The NameFinder system scores reasonably well in both of these categories, an F-Score of 90,43% in domestic news (see table 17) and 91,30% in foreign news (see table 19). The sport news are its weakest point (see table 25), the main reason being sport clubs named after cities. To improve the weaker categories; business news, lifestyle news and sport news separate gazette lists for each might be created. This of course would not help on texts combined of more than one category.

There is still room for lots of improvements, many of which would be very easy to implement, for example, the rule that all proper nouns have to be in the same case, which is too strict for foreign names. Named entities combined of more than one word, where one of the words is recognized as location from a gazette list, might almost certainly be marked as company (Atlanta Hawks, Manchester United). When the first word in a sentence is found to be a first name, it is not learnt as a start of name and therefore not found again if it again appears as the first word of a sentence, not followed by a last name.

Although the NameFinder system is far from perfect it still catches most of the named entities in newspaper text and could therefore be used as a tool to create a large corpus of named entities which might then be used to train a machine learning system. This corpus would of course have to be manually annulated by a person to correct any errors.

7. Acknowledgements

I would like to thank Hrafn Loftsson for his input in this work and the use of *IceTagger.*

8. References

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

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

Bick, E. (2004). A named entity recognizer for Danish. In Lino, M. T. *et al.* (eds), *Proceedings of LREC 2004.*

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

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

Carreras, Xavier; Márques, Lluís; Padró, Lluís. (2002). Named Entity Extraction using AdaBoost In: *Proceedings of CoNLL-2002*, Taipei, Taiwan.

Dalians, H.; Åström, E. (2001). SweNam—a Swedish named entity recognizer. Its construction, training and evaluation. *Technical report, TRITA-NAP0113, IPLab-189, NADA, KTH,* Stockholm: Royal Institute of Technology.

Florian, Radu. (2002). Named Entity Recognition as a House of Cards: Classifier Stacking. In: Proceedings of CoNLL-2002, Taipei, Taiwan.

Klein, G. (2005). JFlex User's Manual. Technical Report 1.4.1. URL http://jflex.de/manual.html.

Já (n.d.) Tungutæknitól – Kenndu þínum lausnum íslensku. Taken from website (May 14 2009) http://ja.is/vorur-og-thjonusta/spurl/tungutaeknitol/

Johannessen, Janne Bondi; Hagen, Kristin; Haaland, Åsne; Jónsdottir, Andra Björk; Kokkinakis, Dimitri; Meurer, Paul; Bick, Eckhard; Haltrup, Dorte (2005). Named Entity Recognition for the Mainland Scandinavian Languages, Literary and Linguistic Computing.

Loftsson , Hrafn. (2008). Tagging Icelandic text: A linguistic rule-based approach. In *Nordic Journal of Linguistics*, 31(1), 47-72. Cambridge University Press.

Loftsson, H.; Rögnvaldsson, E. (2007). IceNLP: A Natural Language Processing Toolkit for Icelandic. In *Proceedings of InterSpeech 2007, Special session: "Speech and language technology for less-resourced languages"*. Antwerp, Belgium.

Loftsson, H. (2007). Tagging Icelandic Text using a Linguistic and a Statistical Tagger. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the ACL*, Rochester, NY, USA.

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

Marsh, Elaine; Perzanowski, Dennis (1998). MUC-7 EVALUATIONOF IE TECHNOLOGY:Overview of Results.

Mikheev, Andrei; Grover, Claire; Moens, Marc .(1998). Description of the LTG system used for MUC-7. In *Proceedings of the Seventh MessageUnderstanding Conference.*

Mikheev, Andrei; Grover, Claire; Moens, Marc . (1999). Named entity recognition without gazetteers. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics.*

Nadeau, David (2007). Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision.

Rannís (2006). Íslenskur textaskimi. Taken from website (May 14 2009) http://www.rannis.is/sjodir/taeknithrounarsjodur/verkefnalisti/nr/1031/

Sekine, Satoshi (1998). NYU: Description of The Japanese NE System Used For Met-2. *Proc. Message Understanding Conference.*

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

Universiteit Antwerpen. (n.d) Language-Independent Named Entity Recognition (I). Taken from website (14.may.2009) http://www.cnts.ua.ac.be/conll2002/ner/

Contents

1. **Introduction** 1

2. **Related Work** 3

   2.1. **Linguistic NER systems** 5

   2.2. **Machine learning NER systems** 6

      2.2.1. **Supervised learning** 6

      2.2.2. **Unsupervised learning** 7

3. **System description** 8

4. **Implementation** 12

5. **Evaluation** 16

6. **Conclusion and future work** 29

7. **Acknowledgements** 29

8. **References** 30