1 Automatic Acquisition of Lexico-semantic Knowledge for Question Answering

1.1 Introduction

Lexico-semantic knowledge is becoming increasingly important within the area of natural language processing, especially for applications, such as Word Sense Disambiguation, Information Extraction and Question Answering (QA). Although the coverage of handmade resources, such as WordNet (Fellbaum, 1998), in general is impressive, coverage problems still exist for those applications involving specific domains or languages other than English.

We are interested in using lexico-semantic knowledge in an open-domain question answering system for Dutch. Obtaining such knowledge from existing resources is possible, but only to a certain extent. The most important resource for our research is the Dutch portion of EuroWordNet (Vossen, 1998), however its size is only half of that of the English WordNet. Therefore, many of the lexical items used in the QA task of the Cross Language Evaluation Forum (CLEF\(^1\)) for Dutch cannot be found in EuroWordNet. In addition, information regarding the classes to which named entities belong, e.g. \textit{Narvik IS-A harbour}, has been shown to be useful for QA, but such information is typically absent from hand-built resources. For these reasons, we are interested in investigating methods which acquire lexico-semantic knowledge automatically from text corpora.

\(^1\text{http://clef-qa.itc.it/}\)
2.1. Automatic Acquisition of Lexico-semantic Knowledge for Question Answering

The remainder of the chapter is organised as follows: In the next section we briefly describe the question types for which lexico-semantic knowledge will be used, and in section 1.3 we describe related work. In section 1.4 we outline our approach to finding words with a distributional similarity. Sections 1.5 and 1.6 detail how the acquired knowledge is employed to improve the performance of our QA system with regards to specific question types, i.e. questions asking for the name of persons which have a specific function in an organisation, e.g. *Who is the secretary general of the UN?*, *WHICH* questions, and definition questions. Finally, the results of an evaluation on three years of CLEF test sets are reported in section 1.7, while Section 1.8 summarizes our conclusions and proposes suggestions for future research.

1.2 Lexico-semantic Knowledge for QA

We will now briefly describe the four question types whose performance we hope to improve using automatically acquired lexical knowledge: *WHICH* questions, definition questions, questions that can be answered by off-line methods, and in particular, function questions.

A question type where the use of lexical knowledge is potentially useful are *WHICH* questions such as:

> Which volcano erupted in June 1991?

A QA system may find various named entities, such as *Philippines* and *Pinatubo*), as potential answers to the question and knowing that *Pinatubo is a volcano* can help to identify the correct answer. Information about named entities is typically absent in hand-made lexical resources. In section 1.6, we describe a method for acquiring such categorised named entities automatically from a parsed corpus.

Definition questions are the second type of question where lexical knowledge proves to be useful:

> Who is Javier Solana?

For CLEF 2005, definition questions were restricted to persons and organisations, with the expectation that answers should provide “some fundamental information” to users who know nothing about the named entity.
Generally speaking, it is hard to determine which information should be used to provide an answer to such questions in general. We tried an approach whereby we used automatically acquired and categorised named entities (NEs) in order to find an appropriate category which is required to be included in the answer. In section 1.6, we describe how this information can be used to determine answers to definition questions.

In addition, off-line methods (Fleischman et al., 2003) can be used to improve the performance of QA system. In off-line QA, plausible answers to highly likely questions are extracted before the actual question has been asked.

For example, if a user asks for the age of a person, the answer is extracted from the table, and it is not necessary to search the source text. Bouma et al. (2005) describe how syntactic patterns are used to extract answers for highly likely question types. A table is constructed for relations, such as the age of a person, the location and date of birth, and the death of a person. To expand the number of facts comprised in these tables we have applied anaphora resolution.

For instance consider the following question:

How old is Ivanisevic?

In order to extract the answer to the above question from the text provided below, it is necessary to analyse it at the discourse level.

*Todd Martin was the opponent of the quiet Ivanisevic. The American, who defeated the local hero Boris Becker a day earlier, was beaten by the 26-year old Croatian during the finals of the Grand Slam Cup [...].*

Among other things, one must correctly identify *Ivanisevic*, located in the first sentence, as the denotation of *the Croatian*, located in the second sentence, in order to extract the correct answer that is stated in the second sentence.

In section 1.6 we explain how automatically acquired and categorised named entities are used to establish that *Ivanisevic is a Croatian*. This information is referred to by the terms 'categorised named entities' or 'instances'.

---

2We use the CLEF-corpus for our experiments. This corpus consists of newspaper text from 1994 and 1995.
Often, questions are asked about the function or role of a particular person:

Who is the chair of Unilever?

These so-called function questions are also answered by the off-line module. The following syntactic pattern could serve as a method of extracting \(\langle \text{Person,Role,Organisation} \rangle\)-tuples from the corpus:

\[
\text{name}(\text{PER}) \xleftarrow{\text{app}} \text{noun} \xrightarrow{\text{mod}} \text{name}(\text{ORG})
\]

Here, the \text{name}(\text{PER}) constituent provides the \text{Person} argument of the relation, the \text{noun} provides the role, and the \text{name}(\text{ORG})-constituent provides the name of the \text{Organisation} and an important source of noise is revealed when this pattern is applied to the parsed corpus in cases where the \text{noun} is not indicating a role or a function:

colleague Henk ten Cate of Go Ahead

Here, the noun \textit{colleague} does not represent a role within the organisation \textit{Go Ahead}.

To remedy this problem, we have collected a list of nouns denoting functions or roles from the Dutch version of the multilingual resource EuroWordNet (EWN) (Vossen, 1998), and restricted the search pattern to nouns occurring in this list:

\[
\text{name}(\text{PER}) \xleftarrow{\text{app}} \text{function} \xrightarrow{\text{mod}} \text{name}(\text{ORG})
\]

While the above modifications help to improve precision, they also hurts recall, as many valid function words present in the corpus are not present in EWN. In section 1.5 we report an experiment in which we have expanded the list of function words extracted from EWN semi-automatically with words of distributional similarity found in the corpus.

### 1.3 Related Work

Syntactic relations have been shown to provide information that can be used to acquire clusters of semantically similar words automatically (Lin,
1.3. Related Work

With a fully parsed version of the Dutch CLEF QA corpus (4.1 million sentences containing a total of 78 million words) at our disposal, we were naturally interested in applying this method to Dutch. In particular, we have followed the strategy of Curran and Moens (2002) who evaluate various similarity measures and weight functions against various thesauri (MacQuarie (Bernard, 1990), Moby (Ward, 1996) and Roget (Roget, 1911)). We implemented the majority of the best performing similarity measures and weights according to the evaluation of Curran and Moens (2002) and evaluated their performance against Dutch EuroWordNet. The results of this experiment are given in section 1.4.

Automatically acquired clusters of semantically similar words can be used to extend and/or enrich existing ontological resources. Alfonseca and Manandhar (2002), for instance, describe a method for expanding WordNet automatically, whereby new concepts are placed in the WordNet hierarchy according to their distributional similarity to words already in the hierarchy. Their algorithm performs a top-down search and stops at the synset that is most similar to the new concept. In section 1.5, we implement a similar technique to expand the class of function words obtained from EuroWordNet.

Pasca (2004) and the team of Pantel and Ravichandran (2004) both present methods for acquiring class labels for instances (categorised named entities) from unstructured text. Pasca (2004) applies lexico-syntactic extraction patterns based on Part-of-Speech tags. Patterns were hand-built initially, and extended automatically by scanning the corpus for the pairs of named entities and classes found with the initial patterns. Patterns which occur frequently in matching sentences could then be added as additional extraction patterns. Pantel and Ravichandran (2004) have proposed an algorithm that takes a list of semantic classes in the form of clusters of words as input. Labels for these clusters are found by looking at four lexico-syntactic relationships apposition (*ayatollah Khomeini*), nominal subject (*Khomeini is an ayatollah*), such as (*Ayatollahs such as Khomeini*), and like (*Ayatollahs like Khomeini*).

In J. Pustejovsky (2002) a different approach is taken. First, a list of semantic subcategories is induced automatically. Second, instances for the induced subtypes are identified from a corpus by using syntactic pattern templates, such as the apposition relation, nominal compounds and definitional constructions.
The above mentioned types of lexico-semantic information have been applied to several QA and QA related tasks.

As described in chapter 17 the Omega ontology, a large terminology ontology resulting from the merging of WordNet (Fellbaum, 1998), Mikrokosmos (Mahesh, 1996; O’Hara et al., 1998) and various other sources have been applied to the multi-lingual QA application called AskCal (Philpot et al., 2002).

As for the use of categorised named entities, Pantel and Ravichandran (2004) conducted two QA experiments - one answering definition questions and the other performing QA information retrieval (IR). These experiments show that both tasks benefit from the use of automatically acquired class labels. Pasca (2004) then applies this information to websearch in order to process list-type queries: SAS, SPSS, Minitab and BMDP are returned in addition to the top documents for the query statistical packages. In J. Pustejovsky (2002) the usability of instances of semantic types in anaphora resolution is discussed for the task of relation extraction. Other researchers have used lexico-semantic information for anaphora resolution (Poesio et al., 2002; Markert et al., 2003). In particular, Markert and Nissim (2005) have extended the corpus-based approach with the introduction of a method that extracts the required knowledge from the Web using shallow lexicosyntactic patterns.

1.4 Extracting semantically similar words

An increasingly popular method for acquiring semantically similar words is to extract words with a distributional similarity from large corpora, the underlying assumption of this approach being that semantically similar words are used in similar contexts. The context of a word $W$ may be defined as the document in which $W$ occurs or the $n$ words surrounding $W$ ($n$-grams, bag of words). Alternatively, the context may be defined syntactically whereby words are found in syntactic relations with other words. The syntactic relation of a target word plus its accompanying words form the context of a target word. Approaches which do not use syntax tend to find more associative relations between words, e.g between patient and hospital, whereas approaches using syntactic context tend to find concepts belonging to the same class, such as doctor and surgeon. As
we are ultimately interested in extending the coverage of a resource such as Dutch EuroWordNet, we have focused on the latter approach.

Most research has been undertaken using a limited number of syntactic relations (Lee, 1999; Weeds, 2003). However, Lin (1998a) shows that a system which uses a wide range of grammatical relations outperforms Hindle’s (1990) results that were based on using information from just the subject and object relation. Apart from the subject and object relation we have used several other grammatical relations: adjective, coordination, apposition and prepositional complement. Examples are given in table 1.1.

1.4.1 Data collection

The Dutch CLEF QA corpus, which consists of 78 million words of Dutch newspaper text (Algemeen Dagblad and NRC Handelsblad ’94/’95) comprises the data used in this experiment. This corpus was parsed automatically using the Alpino parser (Van Noord, 2006) and the result of parsed sentences are dependency graphs that adhere to the guidelines of the Corpus of Spoken Dutch (Moortgat et al., 2000).

From these dependency graphs, we have extracted tuples consisting of the (non-pronominal) head of an NP (either a common noun or a proper name), the dependency relation, and either (1) the head of the dependency relation (for the object, subject, and apposition relation), (2) the head plus a preposition (for NPs occurring inside PPs which are prepositional complements), (3) the head of the dependent (for the adjective and apposition relation) or (4) the head of the other elements of a coordination (for the coordination relation). Examples are given in table 1.1. The number of tuples and the number of non-identical \( \langle \text{Noun, Relation, OtherWord} \rangle \) triples (types) found are given in table 1.2. Note that a single coordination can give rise to various dependency triples, as from a single coordination like \textit{bier, wijn, en noten} (\textit{beer, wine, and nuts}) the triples \( \langle \text{bier, coord, wijn} \rangle, \langle \text{bier, coord, noten} \rangle, \langle \text{wijn, coord, bier} \rangle, \langle \text{wijn, coord, noten} \rangle, \langle \text{noten, coord, bier} \rangle, \text{and} \langle \text{noten, coord, wijn} \rangle \) can be extracted. Similarly, from the apposition \textit{premier Kok} \( \langle \text{premier, hd_app, Kok} \rangle \) and \( \langle \text{Kok, app, premier} \rangle \) are extracted.
8.1. Automatic Acquisition of Lexico-semantic Knowledge for Question Answering

<table>
<thead>
<tr>
<th>grammatical relation</th>
<th>tuples</th>
<th>types</th>
</tr>
</thead>
<tbody>
<tr>
<td>subject</td>
<td>5.639.140</td>
<td>2.122.107</td>
</tr>
<tr>
<td>adjective</td>
<td>3.262.403</td>
<td>1.040.785</td>
</tr>
<tr>
<td>object</td>
<td>2642.356</td>
<td>993.913</td>
</tr>
<tr>
<td>coordination</td>
<td>965.296</td>
<td>2.465.098</td>
</tr>
<tr>
<td>prep. complement</td>
<td>770.631</td>
<td>389.139</td>
</tr>
<tr>
<td>apposition</td>
<td>526.337</td>
<td>602.970</td>
</tr>
</tbody>
</table>

Table 1.2: Number of tuples and non-identical dependency triples (types) extracted per dependency relation.
A vector was built for each noun that was seen at least 10 times in any dependency relation. After applying this cut-off, vectors are present for 83,479 nouns.

1.4.2 Similarity measures and weights

Various vector-based methods can be used to compute the distributional similarity between words. Curran and Moens (2002) report on a large-scale evaluation experiment, where they evaluated the performance of various commonly used methods. Van der Plas and Bouma (2005) present a similar experiment for Dutch, in which they tested most of the best performing measures, as defined by Curran and Moens (2002), where Pointwise Mutual Information (I) and Dice were awarded the best performance rating. We will now explain this weight and similarity measure in further detail.

The information value of a cell in a word vector, which lists how often a word occurs in a specific grammatical relation to a specific word, is not equal for all cells. For instance, a large number of nouns can occur as the subject of the verb *have* whereas only a few nouns may occur as the object of *squeeze*.

Intuitively, the fact that two nouns both occur as subject of *have* tells us less about their semantic similarity than about the fact that two nouns both occur as object of *squeeze*. To account for this intuition, the frequency of occurrence in a vector can be replaced by a weighted score. The weighted score is an indication of the amount of information carried by the particular combination of a noun and its feature (the grammatical relation, and the word heading the grammatical relation). For this experiment we used Pointwise Mutual Information (I) (Church and Hanks, 1989).

\[ I(W, f) = \log \frac{P(W, f)}{P(W)P(f)} \]

where \( W \) is the target word, 
\( P(W) \) is the probability of seeing the word, 
\( P(f) \) is the probability of seeing the feature, 
and \( P(W, f) \) denotes the probability of seeing the word and the feature together.
Dice is a well-known combinatorial measure that computes the ratio between the size of the intersection of two feature sets and the sum of the sizes of the individual feature sets. We used a variant of the Dice-measure, which Curran and Moens (2002) refer to as $\text{Dice}^\dagger$, which incorporates weighted frequency counts:

$$\text{Dice}^\dagger = \frac{2 \sum_f \min(I(W_1, f), I(W_2, f))}{\sum_f I(W_1, f) + I(W_2, f)}$$

where $f$ is the feature, $W_1$ and $W_2$ are the two words that are being compared, and $I$ is a weight assigned to the frequency counts.

1.4.3 Performance

The Dutch version of the multilingual resource EuroWordNet (EWN) (Vossen, 1998) was used for evaluation. 1000 target words from Dutch EWN with a frequency of more than 10 were randomly selected, according to the frequency information present in Dutch EWN. For each word we collected its 100 most similar words (nearest neighbours) in accordance with the system under evaluation, and for each pair of words (target word + one of the most similar words) we calculated the semantic similarity according to Dutch EWN. A system scores well if the nearest neighbours found by the system also have a high semantic similarity according to EWN.

Lin (1998b) evaluates a number of measures for computing WordNet similarity. From the measures, which are defined in terms of is-A relations only, the Wu and Palmer (1994) measure correlated best to human judgement. The Wu/Palmer measure for computing the semantic similarity between two words $W_1$ and $W_2$ in a word net, whose most-specific common ancestor is $W_3$, is defined as follows:
1.5 Using automatically acquired role and function words

<table>
<thead>
<tr>
<th>$k$</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>system</td>
<td>.60</td>
<td>.54</td>
<td>.52</td>
<td>.49</td>
<td>.46</td>
<td>.44</td>
</tr>
</tbody>
</table>

Table 1.3: Average EWN similarity at (k) candidates when combining dependency relations based on Dice†+I

\[
Sim = \frac{2(D3)}{D1 + D2 + 2(D3)}
\]

where $D1$ ($D2$) is the distance from $W1$ ($W2$) to the lowest common ancestor of $W1$ and $W2$: $W3$, and $D3$ is the distance of that ancestor to the root node.

Table 1.3 reports average EWN similarity for the 1, 5, 10, 20, 50, and 100 most similar words for the 1000 words in our test set. If a word is ambiguous according to EWN, i.e. it is a member of several synsets, the highest similarity score is used. The EWN similarity of a set of word pairs is defined as the average of the similarity between the pairs. The baseline for this task is 0.26, which is the score obtained by picking 100 random words to be the nearest neighbours of a given target word. The upper-bound ranges from 0.99 (if only the most similar word is selected) to 0.84 (if 100 similar words have to be selected). van der Plas and Bouma (2005b) show that the system using data obtained from all syntactic relations outperforms systems using only a subset of the syntactic relations. Furthermore, their work reveals that Dice†+I outperforms various other combinations of weight functions and similarity measures.

1.5 Using automatically acquired role and function words

In section 1.2 we explained that for QA we are interested in extracting, off-line, all instances of the following pattern in our corpus:

\[
\text{name(PER)} \xrightarrow{\text{app}} \text{function} \xrightarrow{\text{mod}} \text{name(ORG)}
\]
To obtain a list of words describing a role or function, all words under the node *leider* ‘leader’, numbering 255 in total were extracted from Dutch EWN. The majority of hyperonyms of this node appeared to indicate function words we were interested in, e.g. it contained (the Dutch equivalents of) *king*, *queen*, *president*, *director*, *chair* etc., while other potential candidates, such as *beroep* ‘profession’, seemed less suitable. However, the coverage of this list, when tested on a newspaper corpus, is far from complete. On the one hand, the list contains a fair number of archaic items, while on the other hand, many functions that occur frequently in newspaper text are missing, e.g. Dutch equivalents of *banker*, *boss*, *national team coach*, *captain*, *secretary-general* etc.

To improve recall, we extended the list of function words obtained from EWN semi-automatically with words with a distributional similarity. In particular, for each of the 255 words in the EWN list, we retrieved the 100 words with the most distributional similarity. We gave each retrieved word a score that corresponds to its reverse rank (1st word: 100, 2nd: 99, 3rd: 98 etc.). The overall score for a word is comprised of the sum of the scores it obtained for the individual target words. Thus, words that are semantically similar to several words in the original list will obtain a higher score than words that were returned only once or twice. Words that were already present in the EWN-list were filtered out.

From an informal evaluation of the results obtained, we have learned that many false positives in the expanded list were either named entities or nouns referring to groups of people, e.g. *board*, *committee* etc. The distinction between groups and functions of individuals is difficult to make on the basis of distributional data. For instance, both a *board* and a *director* can make decisions, report results, be criticised, etc. We attempted to filter both proper names and groups automatically, by discarding both noun stems that start with a capital and noun stems which are listed under the node *groep* ‘group’ in EWN.

Finally, we selected the top-1000 of the filtered list and validated it manually. This list contains 644 valid roles or function nouns, which are absent in EWN. A substantial number of the errors that are found in the list are, in fact, nouns which refer to a group, but which are not listed as such in EWN.

The 644 valid nouns were merged with the original EWN list, to form a list of 899 function or role nouns. Next, the relation extraction process was executed using both the original EWN list and the expanded list.
### 1.6 Using automatically acquired categorised NEs

Both Pasca (2004) and Pantel and Ravichandran (2004) describe methods for acquiring labels for named entities from large text corpora and evaluate their results in the context of web search and question answering. Pantel and Ravichandran (2004) use the apposition relation to find potential labels for named entities. As we already had extracted all appositions from the CLEF corpus as part of the vector-based method for finding semantically similar words, we decided to use this information for two other QA tasks as well.

Apart from applying the apposition relation to acquire categorised NEs for which results were given in van der Plas and Bouma (2005a), we used the relation of nominal predicate compliment, e.g. Monica Seles is a tennis player, in which a NE is found with a predicate complement comprising a noun as well.

We extracted 342,180 categorised NE types from 550,198 tokens. 90.6% of the data was obtained using the apposition relation and 9.4% was found by scanning the corpus for predicate complements. For instance, this database contains 391 names of islands (Bali, Bonaire, Aruba etc.) and 186 different queens (Elizabeth, Wilhelmina, Beatrix etc.). The class labels extracted for each named entity may contain a certain amount of noise. However, by focusing on the most frequent label for a named entity, most of the noise can be discarded. For instance, Beatrix occurs 1210 times in

<table>
<thead>
<tr>
<th></th>
<th>EWN</th>
<th>EWN+</th>
</tr>
</thead>
<tbody>
<tr>
<td>tuples</td>
<td>34191</td>
<td>77028</td>
</tr>
<tr>
<td>unique</td>
<td>16530</td>
<td>46589</td>
</tr>
</tbody>
</table>

Table 1.4: Coverage of function table with (EWN+) and without (EWN) expansion

The effect its use has on recall is illustrated in table 1.4. The number of extracted tuples increases by 125%, while the number of types increases by 181%. The effect of this increase on the performance of our QA system is described in section 1.7.
the extracted tuples: 1150 times as queen, and not more than 60 times with various other labels (play, name, hat, possibility etc.).

Regarding the ambiguity of the classified named entities we can say that on average a named entity has 1.9 labels, however, the distribution is skewed: 80% have only one label, while, in an extreme example from the remaining 20% the most ambiguous named entity, the Netherlands, has 704 labels in total. Off-line answer extraction is a technique that tries to extract answers to highly likely question types before the question is posed. We used the extracted class labels to improve the performance of our QA system on WHICH questions such as:

Which ferry sank southeast of the island Utö?

Question analysis and classification tells us that this is a question of type which(ferry). Candidate answers that are selected by our system are: Tallinn, Estonia, Raimo Tiilikainen etc. The QA system uses various strategies to rank potential answers, i.e. the score assigned to the passage by Information Retrieval (IR), the presence of named entities from the question in the sentence in which the answer is found, the syntactic similarity between question and answer sentence, the frequency of the answer in the set of potential answers etc. Still, selecting the correct named entity for answers to WHICH questions poses considerable problems to our system.

To improve the performance of the system regarding these questions, we have incorporated an additional strategy for selecting the correct answer. Potential answers which have been assigned the class corresponding to the question stem (in this case, ferry) are ranked higher than potential answers for which this class label cannot be found in the database of IS-A-relations. Since Estonia is the only potential answer which IS-A ferry, according to our database, this answer is selected. Note that in answering WH-questions we do not just select the most frequent label assigned to a named entity, but rather check as to whether the named entity occurs at least once with the appropriate class label.

A second question type for which the acquired class labels are relevant is definition question. The CLEF 2005 QA test set contains no less than 60 questions of the form:

What is Sabena?

The named entity Sabena occurs frequently in the corpus, but often with class labels assigned to it, which are not suitable for inclusion in a
1.6. Using automatically acquired categorised NEs

definition (possibility, partner, company etc.). By focusing on the most frequent class label assigned to a named entity (airline company in this case), a more appropriate label for a definition can be found. While frequency is important, often the class label by itself is not sufficient for an adequate definition. For this reason, we have expanded the class label with modifiers which typically need to be included in a definition.

In particular, our strategy for answering definition questions consists of two phases:

- Phase 1: The most frequent class found for a named entity is taken.
- Phase 2: The sentences which mention the named entity and the class are selected, and then searched for additional information which might be relevant. Snippets of information that are found in an adjectival relation or a prepositional complement to the class label are selected accordingly.

In the example above, our system produces Belgian airline company as answer.

However, it must be stressed that deciding beforehand what information is relevant is far from being a trivial step. As explained above, we decided to expand only the label with adjectival and PP modifiers that are adjacent to the class label in the corresponding sentence, and this is the reason for a number of answers being inexact. Given the constituent the museum Hermitage in St Petersburg, this strategy fails to include in St Petersburg, for instance. We did not include relative clause modifiers, as these tend to contain information which is not appropriate for a definition. However, in the case of the question, Who is Iqbal Masih?, an answer that includes at least the first conjunct of the relative clause of the constituent twelve year old boy, who fought against child labour and was shot Sunday in his home town Muritke would have been preferable over just selecting twelve year old boy. Similarly, we did not include purpose clauses, which leads the system to respond large scale American attempt, in response to the question What was the Manhattan project?, instead of large scale American attempt to develop the first (that is, before the Germans) atomic bomb.

The extracted class labels were employed in order to be able to extract potential answers from a corpus not only when answers are clearly stated with the accompanying named entity in the same sentence, but also when
an anaphoric expression is used to refer to an earlier mentioned named entity. More specifically we have attempted to resolve the definite NPs, and to find the named entities they refer to. Our strategy is as follows: We scan the left context of the definite NP for named entities from right to left, i.e. the closest named entity is selected first. For each named entity we encounter, we check whether it is in an IS-A relation with the definite NP. If so, the named entity is selected as the antecedent of the NP. As long as no suitable named entity is found we select the previous named entity and so on until we reach the beginning of the document. We have limited our search to the current document. If no suitable named entity is found, i.e., no named entity is found that is in an IS-A relation with the definite NP, then a fallback is used. This fallback comes down so as to extract the NE in the previous sentence: that is nearest to the anaphoric expression, if there is a NE present. If no NE is present in the previous sentence, the NP is not resolved.

If the NP is resolved, this fact is added to the facts table. In order to explain our strategy for resolving definite NPs, we will now apply it to the example from the introduction:

Todd Martin was the opponent of the quiet Ivanisevic. The American who defeated the local hero Boris Becker a day earlier, was beaten by the 26-year old Croatian during the finals of the Grand Slam Cup […].

In the example above, the left context of the NP the 26-year old Croatian is scanned from right to left. The named entity Boris Becker is selected before the correct antecedent Ivanisevic. The fact that Boris Becker is not found in an IS-A relation with Croatian renders it an unsuitable candidate, and it is put aside. Then, Ivanisevic is selected and this candidate is found to be in an IS-A relation with Croatian, so Ivanisevic is taken as the antecedent of Croatian, and the fact Ivanisevic, 26-year old is thus added to the Age table.

1.7 Evaluation

In this section we evaluate the effect of using lexico-semantic knowledge for QA. We discuss the effect of using categorised NEs for anaphora
Table 1.5: Number of facts found for the different tables for relation extraction

<table>
<thead>
<tr>
<th></th>
<th>original</th>
<th>anaphora</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>20.229</td>
<td>24.917</td>
</tr>
<tr>
<td>born_date</td>
<td>2.297</td>
<td>2.395</td>
</tr>
<tr>
<td>born_loc</td>
<td>795</td>
<td>948</td>
</tr>
<tr>
<td>died_age</td>
<td>923</td>
<td>966</td>
</tr>
<tr>
<td>died_date</td>
<td>1.011</td>
<td>1.204</td>
</tr>
<tr>
<td>died_how</td>
<td>1.834</td>
<td>2.336</td>
</tr>
<tr>
<td>died_loc</td>
<td>720</td>
<td>744</td>
</tr>
</tbody>
</table>

Using categorised NEs for anaphora in relation extraction leads to improvements in terms of coverage, as can be seen in table 1.5. The added facts fall into two categories: they are either facts that were already present in the original table or facts that are new. It should be noted that the facts that are not new do contribute to the overall reliability of the table, as facts that are found more frequently are more often correct than facts that are found only once.

We have extracted all differences between entries (types) in the original table and as well in the table that uses anaphora resolution. These differences are either new facts or increases in frequency. From these differences, we then randomly extracted 400 entries. Two of the authors determined the correctness of the found facts in both tables, the results of which are given in table 1.6. A large number of facts (104 from 400) show a rise in frequency and, in fact, 95 of these 104 examples are correct facts. These are positive results, with regard to the reliability of the table. The precision of the facts however is not very encouraging. Overall 263 (66%) of the 400 facts are correct. For the new facts the percentage correct items is even lower (57%). In Mur and Van der Plas (2007) we reported a very high precision for the facts added by using anaphora resolution. There was a difference in precision between the original and the expanded tables of only 1%. The difference with the method used in the current experiment
lies in the fact that we did not use a fallback in the previous study, but rather resolved the anaphora if and only if an IS-A-relation was found between the co-referring NP and the candidate antecedent. Although the precision of the added facts was high, the number of facts found proved to be disappointing, and for this reason, we have chosen to include a fallback in this experiment.

We evaluated the use of lexico-semantic information for QA on the Dutch questions from CLEF ’03, ’04 and ’05. We then compared the performance of two versions of our QA system: the baseline is the version that does not make use of lexico-semantic information, while the improved system uses the different types of lexico-semantic knowledge discussed in this chapter.

This performance of the baseline and improved system is shown in table 1.7. In the first column, the question type is given (question types not relevant to this chapter are left out). In the second and fifth columns the number of questions classified as being of the corresponding question type are shown. In columns 3 and 6, the corresponding mean reciprocal rank (MRR) score is given ³, while in columns 4 and 7 the CLEF score is given. The CLEF score gives the precision of the first (highest ranked) answer only.

The baseline of our QA system was the Joost QA system, without a special question type for function questions, and without access to IS-A-relations. The baseline treats function questions as person questions, i.e. as questions which require a named entity of type person as an answer. WHICH questions and definition questions are answered by selecting the most highly ranked answer from the list of relevant paragraphs returned by

³The MRR score is the average of 1/R where R is the rank of the first correct answer computed over the five highest ranked answers only
1.7. Evaluation

the IR component. Answers to definition questions are in essence selected by means of the same strategy as described for the improved system above, the only exception being that answers must now be selected from the documents returned by IR, rather than from sentences known to contain a relevant class label.

The improved system makes use of the question type function and function_of and the related tables in which information about functions is stored. This last type refers to questions, such as Jeltsin is president of which country? Furthermore, it uses IS-A-relations when answering WHICH questions and definition questions.

The overall effect of adding lexico-semantic information is an improvement in MRR score of 6%, and an improvement in CLEF score of 5%.

Adding a question class for functions and a table for this relation has the effect that 94 person questions and 23 WHICH question in the baseline system are now classified as function or function_of questions. The effect on accuracy of this change appears to be small, as person questions are already answered relatively well, however on the other hand, for WHICH questions, answered with a much lower precision the effect is more noticeable. The shift from person and WHICH questions to function and function_of questions is beneficial. Of the 93 questions that are classified as function or function_of questions in the improved system, 19 involve question stems, such as weduwe ‘widow’, adviseur ‘advisor’, secretaris-generaal ‘secretary-general’, and vriendin ‘girl friend’, which were present in our extended list of function nouns only.

Adding IS-A-relations as an additional knowledge source for answering WH-questions improves the MRR score of 114 WH-questions with 13% and improves the CLEF score with 15%. In addition, using the same information to provide answers to definition questions improves the MRR score of 88 definition questions by 11%, and improves the CLEF score by 13%.

We were not able to show that using lexico-semantic driven anaphora resolution for relation extraction improves the performance of the system on the CLEF test set. We believe that this is due to the fact that the test set contains only 19 questions with a question type for which anaphora resolution potentially could make a difference, i.e. questions that were of one of the question types (see table 1.5) for which the relation extraction module using anaphora resolution provides answers.
1. Automatic Acquisition of Lexico-semantic Knowledge for Question Answering

<table>
<thead>
<tr>
<th>q-type</th>
<th>baseline</th>
<th></th>
<th>improved</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># q</td>
<td>MRR</td>
<td>CLEF</td>
<td># q</td>
</tr>
<tr>
<td>WH</td>
<td>137</td>
<td>0.39</td>
<td>0.33</td>
<td>114</td>
</tr>
<tr>
<td>definition</td>
<td>88</td>
<td>0.47</td>
<td>0.39</td>
<td>88</td>
</tr>
<tr>
<td>person</td>
<td>94</td>
<td>0.72</td>
<td>0.68</td>
<td>30</td>
</tr>
<tr>
<td>funct</td>
<td>0</td>
<td>0.00</td>
<td>0.00</td>
<td>87</td>
</tr>
<tr>
<td>funct_of</td>
<td>0</td>
<td>0.00</td>
<td>0.00</td>
<td>6</td>
</tr>
<tr>
<td>total</td>
<td>773</td>
<td>0.48</td>
<td>0.45</td>
<td>773</td>
</tr>
</tbody>
</table>

Table 1.7: Overall performance of the baseline and improved QA system on the CLEF ('03, '04, '05) Dutch QA test set.

1.8 Conclusions and future work

We have demonstrated that lexico-semantic knowledge can be acquired from syntactically parsed corpora, and that the inclusion of such knowledge in a QA system has a positive effect on the overall performance of the QA system. First, it can be seen that the use of off-line techniques in general has a positive effect on the accuracy of QA. Here, we have demonstrated that the resources required for the off-line extraction of function relations can be acquired semi-automatically, by expanding a given list of relevant function words. Furthermore, coverage of a number of tables can be increased by including anaphora resolution. Second, the performance of the system on WHICH questions and definition questions was shown to improve considerably if it has access to automatically acquired class labels.

The research reported here can be extended and improved upon in several ways. For instance, alternative ways of exploiting the class labels in QA can be explored. Pantel and Ravichandran (2004), for example, use class labels to index the document collection, i.e. every paragraph which mentions a named entity known to be a *ferry* is labelled with this class as well. This strategy allows the IR component to make use of class information. Pantel and Ravichandran (2004) show this modification to improve the precision of IR considerably. In future work, we would like to explore this possibility as well.
1.8. Conclusions and future work

The lexico--semantically driven anaphora resolution can be improved in several ways. The fallback we have used introduced many false positives. We plan to investigate different methods of improving recall without hurting the precision. One possibility is the expansion of the list of categorised named entities by using more corpora and possibly also the Internet as a source. Another possible method is to expand the list by using semantically similar words in the same way as we employed it for expanding the function list.

Acknowledgements

This research was carried out in the project Question Answering using Dependency Relations, which is part of the research program for Interactive Multimedia Information eXtraction, IMIX, financed by NWO, the Dutch Organisation for Scientific Research.
References


References


Roget, P. (1911). Thesaurus of English words and phrases.


