Boosting Question Answering by Deep Entity Recognition

Piotr Przybyła★★

Institute of Computer Science, Polish Academy of Sciences,
Warsaw, Poland,
`P.Przybyla@phd.ipipan.waw.pl`

Abstract. In this paper an open-domain factoid question answering system for Polish, RAFAEL, is presented. The system goes beyond finding an answering sentence; it also extracts a single string, corresponding to the required entity. Herein the focus is placed on different approaches to entity recognition, essential for retrieving information matching question constraints. Apart from traditional approach, including named entity recognition (NER) solutions, a novel technique, called Deep Entity Recognition (DeepER), is introduced and implemented. It allows a comprehensive search of all forms of entity references matching a given WordNet synset (e.g. an impressionist), based on a previously assembled entity library. It has been created by analysing the first sentences of encyclopaedia entries and disambiguation and redirect pages. DeepER also provides automatic evaluation, which makes possible numerous experiments, including over a thousand questions from a quiz TV show answered on the grounds of Polish Wikipedia. The final results of a manual evaluation on a separate question set show that the strength of DeepER approach lies in its ability to answer questions that demand answers beyond the traditional categories of named entities.

1 Introduction

A Question Answering (QA) system is a computer program capable of understanding questions in a natural language, finding answers to them in a knowledge base and providing answers in the same language. So broadly defined task seems very hard; Shapiro (1992) describes it as *AI-Complete*, i.e. equivalent to building a general artificial intelligence. Nonetheless, the field has attracted a lot of attention in Natural Language Processing (NLP) community as it provides a way to employ numerous NLP tools in an exploitable end-user system. It has resulted in valuable contributions within TREC competitions (Dang et al., 2007) and, quite recently, in a system called *IBM Watson* (Ferrucci et al., 2010), successfully competing with humans in the task.

However, the problem remains far from solved. Firstly, solutions designed for English are not always easily transferable to other languages with more complex syntax rules and less resources available, such as Slavonic. Secondly, vast complexity and formidable hardware requirements of *IBM Watson* suggest that there is still a room for improvements, making QA systems smaller and smarter.

This work attempts to contribute in both of the above areas. It introduces RAFAEL (*RApid Factoid Answer Extraction aLgorithm*), a complete QA system for Polish language. It is the first QA system designed to use an open-domain plain-text knowledge base in Polish to address factoid questions not only by providing the most relevant sentence, but also an entity, representing the answer itself. The Polish language, as other Slavonic, features complex inflection and relatively free word order, which poses additional challenges in QA. Chapter 2 contains a detailed description of the system architecture and its constituents.

In the majority of such systems, designers' attention focus on different aspects of a sentence selection procedure. Herein, a different idea is incorporated, concentrating on an entity picking procedure. It allows to compare fewer sentences, likely to contain an answer. To do that, classical Named Entity Recognition (NER) gets replaced by Deep Entity Recognition. *DeepER*, introduced in this work, is a generalisation of NER which, instead of assigning each entity to one of several predefined NE categories, assigns it to a WordNet synset.

For example, let us consider a question: *Which exiled European monarch returned to his country as a prime minister of a republic?*. In the classical approach, we recognise the question as concerning

---

a person and treat all persons found in texts as potential answers. Using DeepER, it is possible to limit the search to persons being monarchs, which results in more accurate answers. In particular, we could utilise information that *Simeon II* (our answer) is a tsar; thanks to WordNet relations we know that it implies being a monarch. DeepER is a generalisation of NER also from another point of view – it goes beyond the classical named entity categories and treats all entities equally. For example, we could answer a question *Which bird migrates from the Arctic to the Antarctic and back every year?*, although *arctic tern* is not recognized as NE by NER systems. Using DeepER, we may mark it as a seabird (hence a bird) and include among possible answers. Chapter 3 outlines this approach.

The entity recognition process requires an *entities library*, containing known entities, their text representations (different ways of textual notation) and WordNet synsets, to which they belong. To obtain this information, the program analyses definitions of entries found in encyclopaedia (in this case the Polish Wikipedia). In previous example, it would use a Wikipedia definition: *The Arctic Tern (Sterna paradisaea) is a seabird of the tern family Sternidae.* This process, involving also redirect and disambiguation pages, is described in section 3.2. Next, having all the entities and their names, it suffices to locate their mentions in a text. The task (section 3.3) is far from trivial because of a complicated named entity inflection in Polish (typical for Slavonic languages, see (Przepiórkowski, 2007)).

DeepER framework provides also another useful service, i.e. automatic evaluation. Usually QA systems are evaluated by verifying accordance between obtained and actual answer based on a human judgement. Plain string-to-string equality is not enough, as many entities have different text representations, e.g. *John F. Kennedy* is as good as *John Fitzgerald Kennedy* and *John Kennedy*, or *JFK* (again, the nominal inflection in Polish complicates the problem even more). However, with DeepER, a candidate answer can undergo the same recognition process and be compared to the actual expected *entity*, not *string*.

Thanks to automatic evaluation vast experiments requiring numerous evaluations may be performed swiftly; saving massive amount of time and human resources. As a test set, authentic questions from a popular Polish quiz TV show[1] are used. Results of experiments, testing (among others) the optimal context length, a number of retrieved documents, a type of entity recognition solution, appear in section 5.1.

To avoid overfitting, the final system evaluation is executed on a separate test set, previously unused in development, and is checked manually. The results are shown in section 5.2 and discussed in chapter 6. Finally, chapter 7 concludes the paper.

2 RAFAEL

As stated in previous chapter, RAFAEL is a computer system solving a task of Polish text-based, open-domain, factoid question answering. It means that provided questions, knowledge base and returned answers are expressed in Polish and may belong to any domain. The system analyses the knowledge base, consisting of a set of plain text documents, and returns answers (as concise as possible, e.g. a person name), supplied with information about supporting sentences and documents.

What are the kinds of requests that fall into the category of *factoid questions*? For the purpose of this study, it is understood to include the following types:

- **Verification** questions, which could be answered by providing a single boolean value (true or false), e.g. *Did Lee Oswald kill John F. Kennedy?*,
- **Option** questions, requiring to select one of available options, e.g. *Which one killed John F. Kennedy: Lance Oswald or Lee Oswald?*,
- **Named entity** questions, expected to be answered by a single named entity, e.g. *When was John F. Kennedy killed?*,
- **Unnamed entity** questions, similar to above, but the expected entity does not belong to traditional named entity types, e.g. *What did Lee Oswald use to kill John F. Kennedy?*,
- **Other name** questions, asking for another name of a given named entity, e.g. *What nickname did John F. Kennedy use during his military service?*,

---

[1] *Jeden z dziesięciu*

– **Multiple named entities** questions, to be answered by a list of named entities, e.g. *Which U.S. presidents were assassinated in office?*.

Although the above list rules out many challenging types of questions, demanding more elaborate answers (e.g. *Why was JFK killed?*, *What is a global warming?*, *How to build a fence?*), it still involves very distinct problems. Although RAFAEL can recognize factoid questions from any of these types and find documents relevant to them (see more in section 2.4 and (Przybyła, 2013b)), its answering capabilities are limited to those requesting single **unnamed entities** and **named entities**. In this document, they are called **entity questions**.

The task description here is similar to the TREC competitions and, completed with test data described in section 4.1, could play a similar role for Polish QA, i.e. provide a possibility to compare different solutions of the same problem. More information about the task, including its motivation, difficulties and a feasibility study for Polish could be found in (Przybyła, 2012).

2.1 Related work

The problem of Question Answering is not new to the Polish NLP community (nor working on other morphologically rich languages), but none of studies presented so far coincides with the notion of plain text-based QA presented above.

First Polish QA attempts date back to 1985, when Vetulani (1988) presented a Polish interface to ORBIS database, containing information about the solar system. The database consisted of a set of PROLOG rules and the role of the system (called POLINT) was to translate Polish questions to appropriate queries. Another early solution, presented by Duclaye et al. (2002), could only work in a restricted domain (business information).

A system dealing with a subset of the TREC tasks was created for Bulgarian by Tanev (2004). His solution answers only three types of questions: Definition, Where-Is and Temporal. He was able to achieve good results with 100 translated TREC questions, using several manually created answer patterns, without NER or any semantic information. Another system for Bulgarian (Simov & Osenova, 2005) participated in the CLEF 2005 competition. Its answer extraction module bases on partial grammars, playing a role of patterns for different types of questions. They could answer correctly 37 of 200 questions, of which only 16 belong to the factoid type. Previously the same team (Osenova et al., 2004) took part in a Bulgarian-English track of the CLEF 2004, in which Bulgarian questions were answered using English texts.

A QA solution was also created for Slovene (Čeh & Ojsteršek, 2009). The task there is to answer students' questions using databases, spreadsheet files and a web service. Therefore, it differs from the problem discussed above by limited domain (issues related to a particular faculty) and the non-textual knowledge base. Unfortunately, no quantitative results are provided in this work.

More recently, several elements of a Polish QA system called *Hipisek* were presented by Walas & Jassem (2010). It bases on a fairly common scheme of transforming a question into a search query and finding the most appropriate sentence, satisfying question constrains. Unfortunately, a very small evaluation set (65 question) and an unspecified knowledge base (gathered by a web crawler) make it difficult to compare the results. In their later works (Walas, 2012; Walas & Jassem, 2011), the team concentrated on spatial reasoning using a knowledge base encoded as a set of predicates.

The approach presented by Piechociński & Mykowiecka (2005) is the closest to the scope of this work, as it includes analysis of Polish Wikipedia content and evaluation is based on questions translated from a TREC competition. Unfortunately, it heavily relies on a structure of Wikipedia entries, making it impossible to use with an arbitrary textual corpus.

A non-standard approach to answer patterns has been proposed by Konopík & Rohlík (2010). In their Czech open-domain QA system they used a set of templates associated with question types, but also presented a method to learn them semi-automatically from search results. Peshterliev & Koychev (2011) in their Bulgarian QA system concentrated on semantic matching between between a question and a possible answer checked using dependency parsing. However, they provide no data regarding an answering precision of the whole system.

The last Polish system worth mentioning has been created by Marcińczuk et al. (2013). Generally, their task, called Open Domain Question Answering (ODQA), resembles what is treated here, but with one major difference. A document is considered an answer; therefore they focus on

improving ranking in a document retrieval stage. They have found out that it could benefit from taking nearness of query terms occurrences into account.

As some of Slavonic languages lack necessary linguistic tools and resources, only partial solutions of QA problems exist for them, e.g. document retrieval for Macedonian (Armenska et al., 2010), question classification for Croatian (Lombarović et al., 2011) or answer validation for Russian (Solovyev, 2013).

2.2 System Architecture

A general architectural scheme of RAFAEL (figure 1) has been inspired by similar systems developed for English; for examples see works by Hovy et al. (2000) and Moldovan et al. (2000).
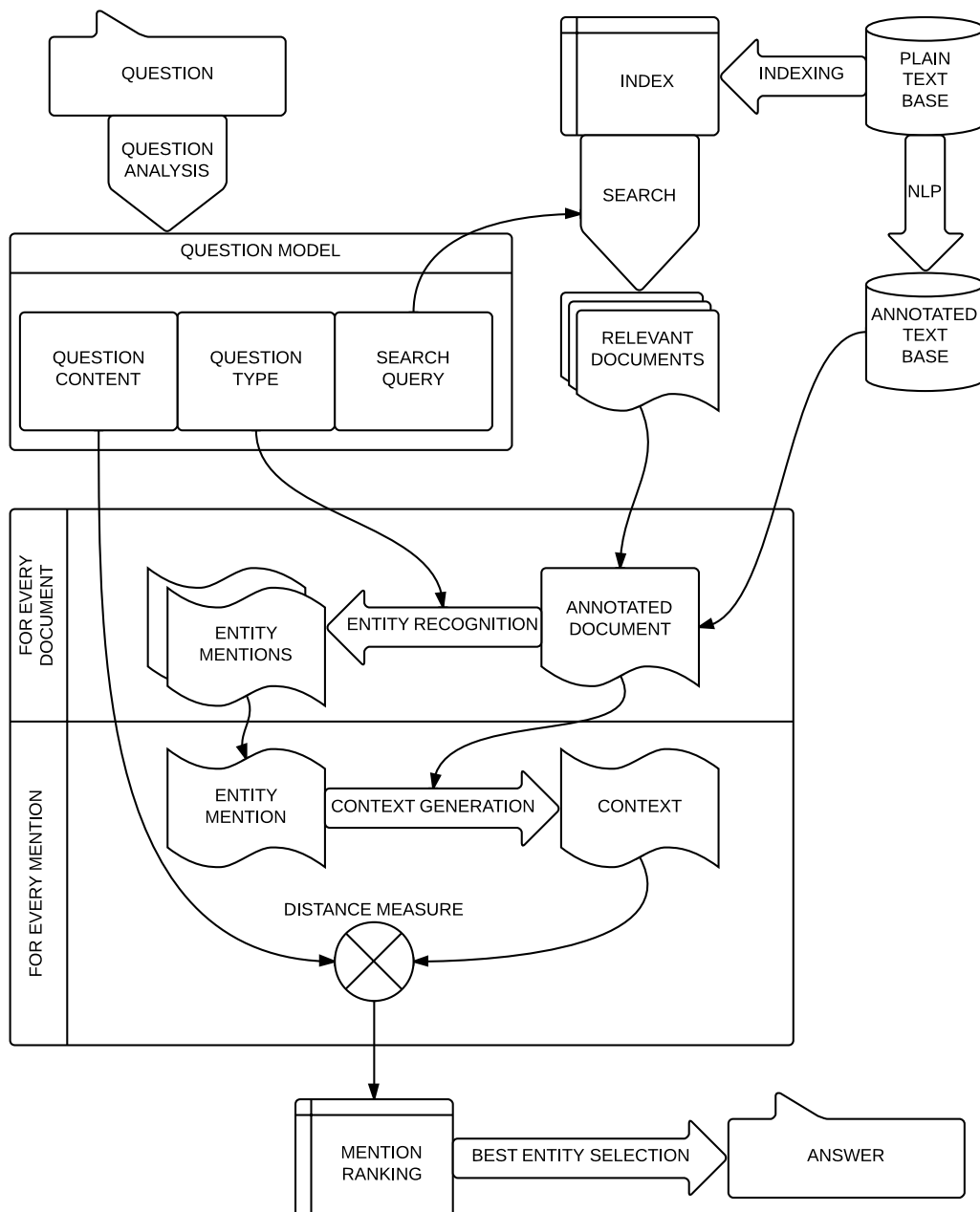
Fig. 1. Overall architecture of the QA system – RAFAEL. See descriptions of elements in text.

Two of the steps in the diagram concern offline processing of a knowledge base. Firstly, it is indexed by a search engine to ensure efficient searching in further stages (INDEXING). Secondly, it may be annotated using a set of tools (NLP), but this could also happen at an answering stage for selected documents only.

After the system receives a question, it gets analysed (QUESTION ANALYSIS) and transformed into a data structure, called *question model*. One of its constituents, a search query, is used to find a set of documents, which are probably appropriate for the current problem (SEARCH). For each of the documents, all entity mentions compatible with an obtained question type (e.g. monarchs), are extracted (ENTITY RECOGNITION). For each of them, a context is generated (CONTEXT GENERATION). Finally, a distance between a question content and the entity context is computed to asses its relevance (DISTANCE MEASURE). All the mentions and their distance scores are stored and, after no more documents are left, used to select the best match (BEST ENTITY SELECTION). The system returns the entity, supplied with information about a supporting sentence and a document, as an answer.

2.3 Knowledge Base Processing

Knowledge base (KB) processing consists of two elements: indexing and annotating. The objective of the first is to create an index for efficient searching using a search engine. In the system, *Lucene* 3.6[2] is used to build two separate full-text indices: regular and stemmed using a built-in stemmer for Polish, *Stempel* (Galambos, 2001).

Secondly, texts go through a cascade of annotation tools, enriching it with the following information:

- Morphosyntactic interpretations (sets of tags), using *Morfeusz* 0.82 (Woliński, 2006),
- Tagging (selection of the most probable interpretation), using a transformation-based learning tagger, *PANTERA* 0.9.1 (Acedański, 2010),
- Syntactic groups (possibly nested) with syntactic and semantic heads, using a rule-based shallow parser *Spejd* 1.3.7 (Przepiórkowski, 2008) with a Polish grammar, including improved version of modifications by Degórski (2012), enabling lemmatisation of nominal syntactic groups,
- Named entities, using two available tools: *NERF* 0.1 (Savary & Waszczuk, 2012) and *Liner2* 2.3 (Marcińczuk & Janicki, 2012).

All the annotations are stored in a variant of *TEI P5* standard, designed for the National Corpus of Polish (Przepiórkowski et al., 2012). As noted previously, the process of annotating is not indispensable at the stage of offline KB processing; it could be as well executed only on documents returned from the search engine (for example see *Webclopedia* by Hovy et al. (2000) or *LASSO* by Moldovan et al. (2000)). However, since during evaluation experiments the same documents undergo the process hundreds of times, it seems reasonable to process the whole KB only once.

2.4 Question Analysis

The goal of question analysis is to examine a question and extract all the information that suffices for answer finding. A resulting data structure, called *question model*, contains the following elements:

1. *Question type* – a description of expected answer type, instructing the system, what type of data could be returned as an answer. It has three levels of specificity:
   (a) *General question type* – one of the types of factoid questions, enumerated at the beginning of this chapter,
   (b) *Named entity type* – applicable only in case general type equals *named entity*. Possible values are the following: place, continent, river, lake, mountain, mountain range, island, archipelago, sea, celestial body, country, state, city, nationality, person, first name, last name, band, dynasty, organisation, company, event, date, century, year, period, number, quantity, vehicle, animal, title.

---

[2] Available from `http://lucene.apache.org/`.

(c) *Focus synset* – applicable in case of entity questions; a WordNet synset, to which a question focus belongs; necessary for DeepER.
2. *Search query* – used to find possibly relevant documents,
3. *Question content* – the words from question which are supposed to appear also in context of an answer.

The task presented above, called *question classification*, is an example of text classification with very short texts. It could be tackled by a general-purpose classifier; for example, Čeh & Ojsteršek (2009) used SVMs (Support Vector Machines) for closed-domain Slovene QA system; Li & Roth (2002) employed SNoW (Sparse Network of Winnows) for hierarchical classification of TREC questions. For Polish results are not satisfactory (Przybyła, 2013b) because of data sparsity.

However, sometimes a solution seems quite evident, as part of the question types enforce its structure. For example, when it begins with *Who* or *When*, it belongs to *person* and *date* question types, respectively. That is why a set of 176 regular expressions (in case of RAFAEL) suffices to deal with them. They match only a subset of questions (36.15 per cent of the training set), but are highly unambiguous (precision of classification equals 95.37 per cent). Nevertheless, some (Lee et al., 2005) use solely such patterns, but need a great number of them (1,273).
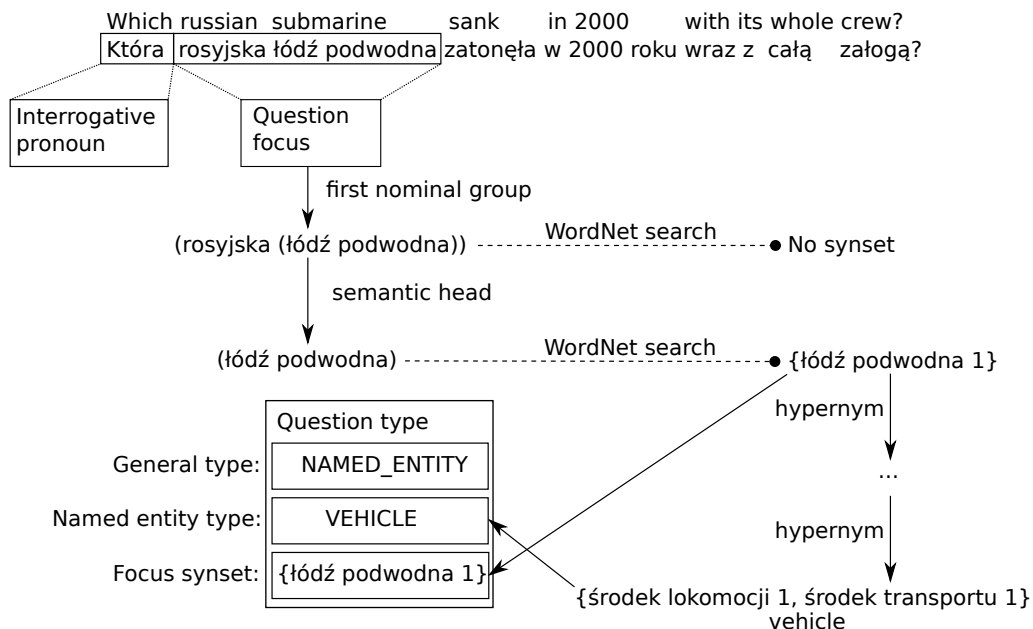

Fig. 2. Outline of a question focus analysis procedure used to determine an entity type in case of ambiguous interrogative pronouns.

Unfortunately, most of entity questions are ambiguous, i.e. it is not enough to inspect an interrogative pronoun to find an answer type. They may begin with *what* or *which*, followed by a *question focus*. For example, let us consider a question *Which russian submarine sank in 2000 with its whole crew?*. Its focus (*russian submarine*) carries information that the question could be answered by a named entity of type *vehicle*. The whole process of focus analysis is shown in figure 2. The first nominal group after a pronoun[3] serves as a possible lexeme name in *plWordNet* 2.1 (Maziarz et al., 2012). As long as there are no results, it gets replaced by its semantic head. When a matching lexeme exists in WordNet, a set of all its hypernyms is extracted. If any of the elements in the set correspond to one of the named entity types, this type is recorded in the question model. Otherwise the general question type takes the value *unnamed entity*. A WordNet-assisted focus analysis was also implemented in one of solutions participating in a TREC competition (Harabagiu et al., 2001).

---

[3] RAFAEL includes a manually created list of *opening constructions* to be ignored if appearing right after a pronoun, such as *typ* (type of) or *spośród* (out of).

Search query generation is described in the next chapter. The last element of a question model, called *question content*, contains segments, which are to be compared with texts to find the best answer. It includes all the words of the interrogative sentence except those included in the matched pattern (*Which, ?*) and the focus (*submarine*). In our example the following are left: *russian, sank, in, 2000, with, its, whole, crew*. An entity mention, which context resembles this set, will be selected as an answer (see details in section 2.7).

The question analysis stage explained above follows a design presented in previous works (Przybyła, 2013b,a), where more details could be found. The major difference lies in result processing – an original synset is not only projected to one of the named entity types, but also recorded as a *focus synset* in question type, utilised in DeepER to match entity types. In our example, it would only consider submarines as candidate answers.

2.5 Document Retrieval

The use of search engines in QA systems is motivated mainly by performance reasons. Theoretically, we could analyse every document in a text base and find the most relevant to our query. However, it would take excessive amount of time to process the documents, majority of which belong to irrelevant domains (839,269 articles in the test set). A search engine is used to speed up the process by selecting a set of documents and limiting any further analysis to them.

As described in section 2.3, a knowledge base is indexed by *Lucene* offline. Given a question, we need to create a search query. The problem is that an answer in the knowledge base is probably expressed differently than the question. Hence, a query created directly from words of the question would not yield results, unless using a highly-redundant KB, such as the WWW (for this type of solution see (Brill et al., 2002)). Therefore, some of the query terms should be dropped – based on their low IDF (Katz et al., 2003) or more complex heuristics (Moldovan et al., 2000). On the other hand, the query may be expanded with synonyms (Hovy et al., 2000) or derived morphological forms (Katz et al., 2003).

Finally, we need to address term matching issue – how to compare a query keyword and a text word in a morphologically-rich language, such as Polish? Apart from exact match, it also is possible to use a stemmer or fuzzy queries, available in *Lucene* (accepting a predefined Levenshtein distance between matching strings).

Previous experiments (Przybyła, 2013a) led to the following query generation procedure:

1. Remove all words matched by a regular expression at the classification stage (*What, Which*, etc.),
2. Keep a question focus,
3. Connect all the remaining words by OR operator,
4. Use fuzzy term matching strategy with absolute distance equal 3 characters and fixed prefix.

Lucene handles a query and yields a ranked document list, of which N first get transferred to further analysis. The influence of value of N on answering performance is evaluated in section 5.1.

2.6 Entity Recognition

Having a set of proposed documents and a question type, the next step is to scan them and find all mentions of entities with appropriate types. RAFAEL includes two approaches to the problem: classical Named Entity Recognition (NER) and novel Deep Entity Recognition.

Three NERs for Polish are employed: *NERF, Liner2* and *Quant. NERF* (Savary & Waszczuk, 2012) is a tool designed within the project of the National Corpus of Polish and bases on linear-chain conditional random fields (CRF). It recognizes 13 types of NEs, possibly nested (e.g. *Warsaw* in *University of Warsaw*). *Liner2* (Marcińczuk & Janicki, 2012) also employs CRFs, but differentiates NEs of 56 types (which could be reduced to 5 for higher precision). Annotation using both of the tools happens offline within the KB preprocessing, so in the currently described stage it suffices to browse the annotations and find matching entities. As the above tools lack recognition of quantitative expressions, a new one has been developed especially for RAFAEL and called *Quant*. It is able to handle both numbers and quantities (using *WordNet*) in a variety of notations.

Appendix A contains details of implementation of named entity recognition in RAFAEL, including a description of *Quant* and a mapping between question types and named entity types

available in *NERF* and *Liner2*. An alternative being in focus of this work, i.e. DeepER approach, is thoroughly discussed in chapter 3.

RAFAEL may use any of the two approaches to entity recognition: NER (via *NERF*, *Liner2* and *Quant*) or novel DeepER; this choice affects its overall performance. Experiments showing precision and recall of the whole system with respect to applied entity recognition technique are demonstrated in section 5.1.

2.7   Mention selection

When a list of entity mentions in a given document is available, we need to decide which of them most likely answers the question. The obvious way to do that is to compare surroundings of every mention with the content of the question. The procedure consists of two steps: context generation and similarity measurement.

Context generation The aim of a context generation step is to create a set of segments surrounding an entity, to which they are assigned. Without capabilities of full text understanding, two approximate approaches seem legitimate:

 – Sentence-based – for a given entity mention, a sentence in which it appears, serves as a context,
 – Segment-based – for a given entity mention, every segment sequence of length *M*, containing the entity, is a context.

Both of them have some advantages: relying on a single sentence ensures relation between an entity and a context, whereas the latter provides possibility of modifying context length. Obviously, the value of *M* should be proportional to question (precisely, its content) length.

The method of treating sentences as a context has gained most popularity (see work of Yih et al. (2013)), but a window of fixed size also appears in the literature; for example Katz et al. (2003) used one with M=140 bytes.

The context generation is also related to another issue, i.e. anaphoric expressions. Some segments (e.g. *this*, *him*, *they*) may refer to entities that occurred earlier in a text and therefore harm a similarity estimation. It could be tackled by applying anaphora resolution, but a solution for Polish (Kopeć & Ogrodniczuk, 2012) remains in an early stage. Observations show that the majority of anaphora refer to an entity in a document title, so the problem is partially bypassed by adding a title to a context.

An influence of the context generation techniques on final results is shown in section 5.1.

Similarity measurement To measure a similarity between a question content (explained in section 2.4) and an entity context (generated by the procedures in previous section), a Jaccard similarity index (Jaccard, 1901) is computed. However, not all word co-occurrences matter equally (e.g. compare *this* and *Honolulu*), so word weights are used:

$$\mathrm{Sim}_w(A, B) = \frac{\sum_{i \in A \cap B} w_i}{\sum_{i \in A \cup B} w_i}$$

The sets A and B contain segments in base forms, whereas w_i denotes a weight of an i-th base form, equal to its scaled IDF computed on a document set D:

$$w_i = \frac{\log \frac{|D|}{|\{d:i \in d\}|}}{\max_i \log \frac{|D|}{|\{d:i \in d\}|}}$$

The Jaccard index is a popular solution for sentence similarity measurement in QA (for example see a system by Ahn et al. (2004)). In case of selecting relevant documents, cosine measure is also applied. Marcińczuk et al. (2013) compared it to Minimal Span Weighting (MSW) and observed that the latter performs better, as it takes into account a distance between matched words. A study of different techniques for sentence similarity assessment could be found in (Yih et al., 2013).

Answer generation At this stage, a large set of pairs of entity mention and its contexts with scores assigned, is available. Which of them answers the question? Choosing the one with the highest score seems an obvious solution, but we could also aggregate scores of different mentions corresponding to the same answer (entity), e.g. compute their sum or mean. However, such experiments did not yield improvement, so RAFAEL returns only a single answer with the highest score.

An answer consists of the following elements: an answer string, a supporting sentence, a supporting document and a confidence value (the score). A sentence and a document, in which the best mention appeared, are assumed to support the answer. Thanks to properties of Jaccard similarity, the mention score ranges between 0 for completely unrelated sentences to 1 for practically (ignoring inflection and a word order) the same. Therefore, it may serve as an answer confidence.

When no entity mentions satisfying constraints of a question are found, no answer is returned. This type of result could also be used when the best confidence score is below a predefined value; performance of such technique are shown in section 5.1. The refusal to answer in case of insufficient confidence plays an important role in *Jeopardy!*, hence in *IBM Watson* (Ferrucci et al., 2010), but it was also used to improve precision in other QA systems (Oh et al., 2013).

3 Deep Entity Recognition

Deep Entity Recognition procedure is an alternative to applying Named Entity Recognition in QA to find entities matching question constraints. It scans a text and finds words and multi-word expressions, corresponding to entities. However, it does not assign them to one of several NE categories; instead, WordNet synsets are used. Therefore, named entities are differentiated more precisely (e.g. monarchs and athletes) and entities beyond the classical NE categories (e.g. species, events, devices) could also be recognised in a text.

It does not seem possible to perform this task relying solely on features extracted from words and surrounding text (as in NER), so it is essential to build an entity library. Such libraries already exist (*Freebase*, *BabelNet*, *DBpedia* or *YAGO*) and could provide an alternative for DeepER, but they concentrate on English. The task of adaptation of such a base to another language is far from trivial, especially for Slavonic languages with complex NE inflection (Przepiórkowski, 2007). An ontology taking into account Polish inflection (*Prolexbase*) has been created by Savary et al. (2013), but it contains only 40,000 names, grouped into 34 types.

3.1 Related work

The idea of DeepER in a nutshell is to improve QA by annotating a text with WordNet synsets using an entity base created by understanding definitions found in encyclopaedia. Parts of this concept have already appeared in the NLP community.

A technique of coordinating synsets assigned to question and a possible answer emerged in a study by Mann (2002). While a question analysis there seems very similar to this work, entity library (called *proper noun ontology*) generation differs a lot. The author analysed 1 GB of newswire text and extracted certain expressions, e.g. *"X, such as Y"* implies that Y is an instance of X. Albeit precision of resulting base was not very good (47 per cent for non-people proper names), it led to a substantial improvement of QA performance.

The idea of analysing encyclopaedic definitions to obtain this type of information already appeared, but was employed for different applications. For example, Toral & Muñoz (2006) described a method of building a gazetteer by analysing hyperonymy branches of nouns of first sentences in Wikipedia definitions. Unlike in this work, an original synset was replaced by a coarse-grained NER category. Another example of application is a NE recognizer (Kazama & Torisawa, 2007) using words from a definition as additional features for a standard CRF classifier. In their definition analysis only the last word of the first nominal group was used.

Other researchers dealt with a task explicitly defined as classifying Wikipedia entries to NER categories. For example Dakka & Cucerzan (2008) addressed the problem by combining traditional text classification techniques (bag of words) with contexts of entity mentions. Others (Ponzetto & Strube, 2007) thoroughly examined article categories as a potential source of is-a relations in a taxonomy (99 per cent of entries have at least one category). Inhomogeneity of categories turned

out as the main problem, dealt with by a heuristic classifier, assigning is-a and not-is-a labels. Categories were also used as features in a NER task (Richman & Schone, 2008), but it required a set of manually designed patterns to differentiate between categories of different nature.

Exploring a correspondence between Wikipedia entries and WordNet synsets found an application in automatic enriching ontologies with encyclopaedic descriptions (Ruiz-Casado et al., 2005). However, only NEs already appearing in the WordNet were considered. The task (solved by bag-of-words similarity) is non-trivial only in case of polysemous words, e.g. which of the meanings of *Jupiter* corresponds to which Wikipedia article? Others (Toral et al., 2008) concentrated on the opposite, i.e. extending the WordNet by NEs that are *not* there yet by adding titles of entries as instances of synsets corresponding to their common category.

Also, some see Wikipedia as an excellent source of high-quality NER training data. Again, it requires to project entries to NE categories. A thorough study of this problem, presented by Balasuriya et al. (2009), utilizes features extracted from article content (bag of words), categories, keywords, inter-article and inter-language links. A final annotated corpus turns out as good for NER training as a manually annotated gold standard.

Finally, some researchers try to generalise NER to other categories, but keep the same machine-learning-based approach. For example, Ciaramita & Altun (2006) developed a tagger, assigning words in a text to one of 41 *supersenses*. Supersenses include NE categories, but also other labels, such as *plant*, *animal* or *shape*. The authors projected word-sense annotations of publicly available corpora to supersenses and applied perceptron-trained Hidden Markov Model for sequence classification, obtaining precision and recall around 77 per cent.

3.2 Entity Library

An entity library for DeepER contains knowledge about entities that is necessary for deep entity recognition. Each of them consists of the following elements (entity #9751, describing the Polish president, Bronisław Komorowski):

- Main name: *Bronisław Komorowski*,
- Other names (aliases): *Bronisław Maria Komorowski*, *Komorowski*,
- Description URL: `http://pl.wikipedia.org/wiki/?curid=121267`,
- plWordNet synsets:
    - *<podsekretarz1, podsekretarz stanu1, wiceminister1>* (vice-minister, undersecretary),
    - *<wicemarszałek1>* (vice-speaker of the Sejm, the Polish parliament),
    - *<polityk1>* (politician),
    - *<wysłannik1, poseł1, posłaniec2, wysłaniec1, posłannik1>* (member of a parliament),
    - *<marszałek1>* (speaker of the Sejm),
    - *<historyk1>* (historian),
    - *<minister1>* (minister),
    - *<prezydent1, prezydent miasta1>* (president of a city, mayor).

A process of entity library extraction is performed offline, before question answering. The library built for deep entity recognition in RAFAEL, based on the Polish Wikipedia (857,952 articles, 51,866 disambiguation pages and 304,823 redirections), contains 809,786 entities with 1,169,452 names (972,592 unique). The algorithm does not depend on any particular feature of Wikipedia, so any corpus containing entity definitions could be used.

Figure 3 shows an exemplary process of converting the first paragraph of a Polish Wikipedia entry, describing former Polish president Lech Wałęsa[4], into a list of WordNet synsets. First, we omit all unessential parts of the paragraph (1). This includes text in brackets or quotes, but also introductory expressions like *jeden z* (one of) or *typ* (type of). Then, an entity name is detached from the text by matching one of definition patterns (2). In the example we can see the most common one, a dash (–). Next, all occurrences of separators (full stops, commas and semicolons) are used to divide the text into separate chunks (3). The following step employs shallow parsing annotation – only nominal groups that appear at the beginning of the chunks are passed on (4). The first chunk that does not fulfil this requirement and all its successors get excluded from further analysis (4.1). Finally, we split the coordination groups and check, whether their lemmas correspond

---

[4] `http://pl.wikipedia.org/wiki/Lech_Wa%C5%82%C4%99sa`

to any lexemes in WordNet (5). If not, the process repeats with the group replaced by its semantic head. In case of polysemous words, only the first word sense (usually the most common) is taken into account.

The whole process is more complicated than the simple example shows. Generally, it consists of the following steps:

Step 0 Prepare a corpus – data format and annotation process is the same as for a knowledge base, used in question answering, see section 2.3. It differs in scope of page categories, including not only articles, but also disambiguation and redirection pages.

Step 1 For each of article pages, extract the first paragraph and apply *readDefinition* function. If a resulting entity has a non-empty synset list, add it to the library. If some of the redirection pages point to the entity name, add their names as entity aliases.

Step 2 For each of disambiguation pages, extract all items and apply *readDefinition* function. If an item refers to an existing entity, extend it with extracted synsets and disambiguation page name. Create a new entity otherwise. Add redirection names as previously.

Step 3 Save the obtained base for future use.

Input: *text* - annotated first paragraph of an encyclopaedic entry
Output: *synsets* - synsets describing an entity
begin
    synsets := {};
    text := removeInBrackets(*text*);
    text := removeInQuotes(*text*);
    foreach *pattern* **in** *definitionPatterns* **do**
        if *pattern matches text* **then**
            definition := match(*pattern*,*text*).group(2);
            break;

    definition := removeDefinitionPrefixes(*definition*);
    parts := split(*definition*,*seperators*);
    foreach *part* **in** *parts* **do**
        chunk := firstGroupOrWord(*part*);
        if *isNominal(chunk)* **then**
            synsets := *synsets* ∪ extractSynsets(*chunk*);
        else break;
    return *synsets*;

Algorithm 1: Function **readDefinition**(*text*) – interprets a definition to assign synsets to an entity.

The *readDefinition* function (shown as algorithm 1) analyses a given paragraph of text and extracts a set of synsets, describing an entity, to which it corresponds, as exemplified by figure 3. Simplifying, it is done by removing all unnecessary text (in brackets or quotes), splitting it on predefined separators (commas, full stops, semicolons) and applying *extractSynsets* function with an appropriate stop criterion. The *readDefinition* makes use of the following elements:

removeInBrackets() removes everything that is between brackets ([], () or {}) from the text (step (1) in figure 3).

removeInQuotes() removes everything between single or double quotes from the text (step (1) in the example).

definitionPatterns contains patterns of strings separating a defined concept from a definition, e.g. hyphens or dashes (used in step (2) of the example) or *jest to* (is a).

removeDefinitionPrefixes() removes expressions commonly prefixing a nominal group, such as *jeden z* (one of), *typ* (a type of) or *klasa* (a class of), not present in the example.

separators a set of three characters that separate parts of a definition: ".", "," and ";".

firstGroupOrWord() returns the longest syntactic element (syntactic group or word) starting at the beginning of a chunk (step (4) in the example).
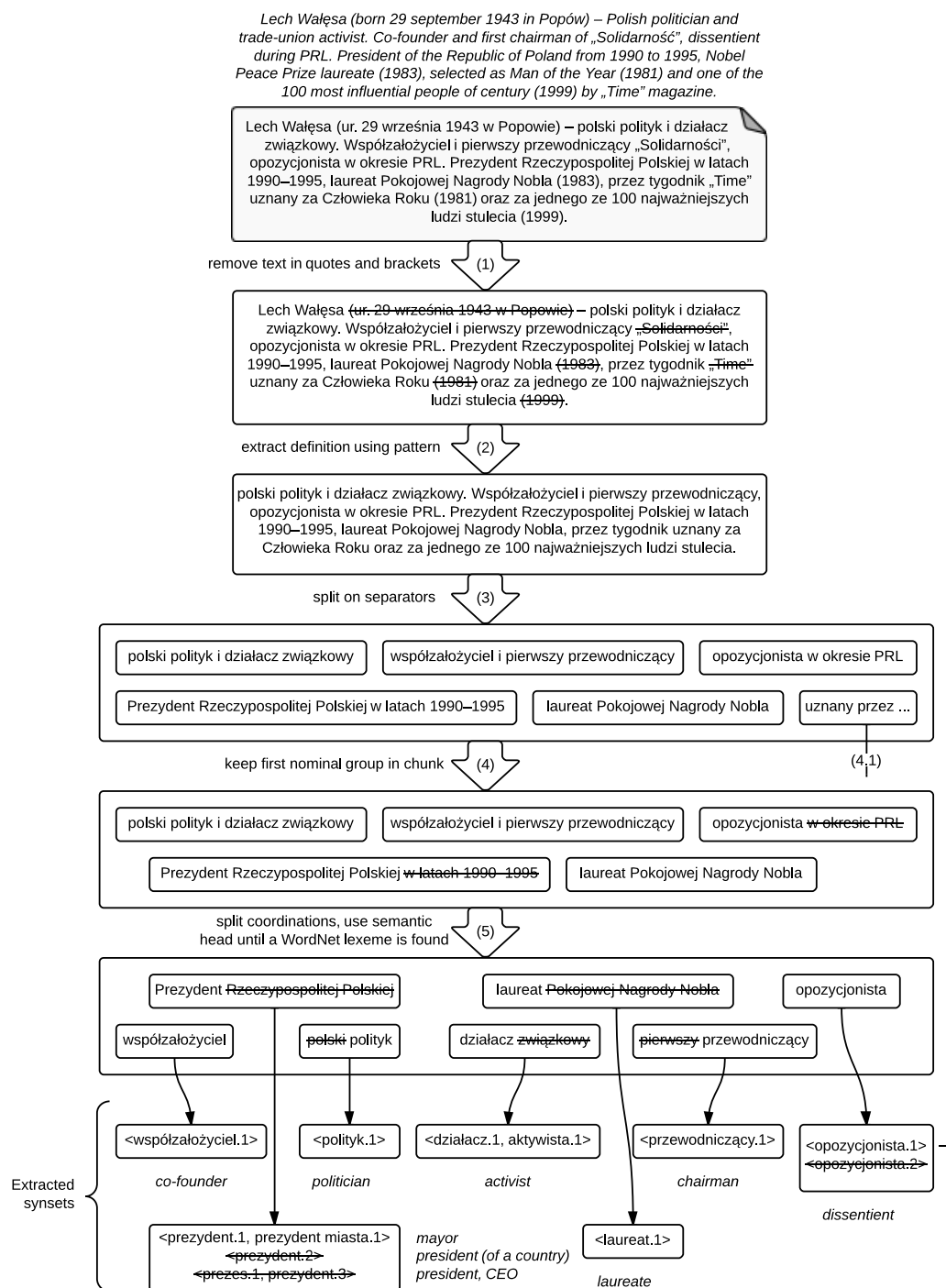
Lech Wałęsa (born 29 september 1943 in Popów) – Polish politician and trade-union activist. Co-founder and first chairman of „Solidarność", dissentient during PRL. President of the Republic of Poland from 1990 to 1995, Nobel Peace Prize laureate (1983), selected as Man of the Year (1981) and one of the 100 most influential people of century (1999) by „Time" magazine.

Lech Wałęsa (ur. 29 września 1943 w Popowie) – polski polityk i działacz związkowy. Współzałożyciel i pierwszy przewodniczący „Solidarności", opozycjonista w okresie PRL. Prezydent Rzeczypospolitej Polskiej w latach 1990–1995, laureat Pokojowej Nagrody Nobla (1983), przez tygodnik „Time" uznany za Człowieka Roku (1981) oraz za jednego ze 100 najważniejszych ludzi stulecia (1999).

remove text in quotes and brackets (1)

Lech Wałęsa ~~(ur. 29 września 1943 w Popowie)~~ – polski polityk i działacz związkowy. Współzałożyciel i pierwszy przewodniczący ~~„Solidarności"~~, opozycjonista w okresie PRL. Prezydent Rzeczypospolitej Polskiej w latach 1990–1995, laureat Pokojowej Nagrody Nobla ~~(1983)~~, przez tygodnik ~~„Time"~~ uznany za Człowieka Roku ~~(1981)~~ oraz za jednego ze 100 najważniejszych ludzi stulecia ~~(1999)~~.

extract definition using pattern (2)

polski polityk i działacz związkowy. Współzałożyciel i pierwszy przewodniczący, opozycjonista w okresie PRL. Prezydent Rzeczypospolitej Polskiej w latach 1990–1995, laureat Pokojowej Nagrody Nobla, przez tygodnik uznany za Człowieka Roku oraz za jednego ze 100 najważniejszych ludzi stulecia.

split on separators (3)

polski polityk i działacz związkowy	współzałożyciel i pierwszy przewodniczący	opozycjonista w okresie PRL
Prezydent Rzeczypospolitej Polskiej w latach 1990–1995	laureat Pokojowej Nagrody Nobla	uznany przez ...

keep first nominal group in chunk (4)   (4,1)

polski polityk i działacz związkowy	współzałożyciel i pierwszy przewodniczący	opozycjonista ~~w okresie PRL~~
Prezydent Rzeczypospolitej Polskiej ~~w latach 1990–1995~~	laureat Pokojowej Nagrody Nobla	

split coordinations, use semantic head until a WordNet lexeme is found (5)

Prezydent ~~Rzeczypospolitej Polskiej~~   laureat ~~Pokojowej Nagrody Nobla~~   opozycjonista

współzałożyciel   ~~polski~~ polityk   działacz ~~związkowy~~   ~~pierwszy~~ przewodniczący

Extracted synsets

<współzałożyciel.1>   <polityk.1>   <działacz.1, aktywista.1>   <przewodniczący.1>   <opozycjonista.1> ~~<opozycjonista.2>~~

co-founder   *politician*   *activist*   *chairman*   *dissentient*

<prezydent.1, prezydent miasta.1> ~~<prezydent.2>~~ ~~<prezes.1, prezydent.3>~~   *mayor president (of a country) president, CEO*   <laureat.1>   *laureate*

Fig. 3. Example of the entity extraction process in DeepER, transforming a Wikipedia entry of *Lech Wałęsa* into a list of synsets.

isNominal() decides, whether a chunk is a noun in nominative, a nominal group or a coordination of nominal groups.

Input: *chunk* - a nominal chunk (a syntactic group or a single noun)
Output: *synsets* - WordNet synsets corresponding to *chunk*
begin
    lemma := lemmatise(*chunk*);
    if *inWordNet(lemma)* **then**
        return *getLexemes(lemma).synset(0)*;
    else if *isCoordination(chunk)* **then**
        synsets := {};
        foreach *element* **in** *chunk* **do**
            synsets := *synsets* $\cup$ extractSynsets(*element*);
        return *synsets*;
    else if *isGroup(chunk)* **then**
        return *extractSynsets(chunk.semanticHead)*;
    else return *{}*;

Algorithm 2: Function **extractSynsets**(*chunk*) – recursively extracts synsets from a nominal chunk.

The *extractSynsets* function (shown as algorithm 2) accepts a nominal chunk and extracts WordNet synsets, corresponding to it. It operates recursively to dispose any unnecessary chunk elements and find the longest subgroup, having a counterpart in WordNet. It corresponds to step (5) in figure 3 and uses the following elements:

lemmatise() returns a lemma of a nominal group.
inWordNet() checks whether a given text corresponds to a lexeme in WordNet.
getLexemes() return a list of WordNet lexemes corresponding to a given text.
synset() return a synset including a lexeme in a given word sense number.
isCoordination() return TRUE iff a given chunk is a coordination group.
isGroup() return TRUE iff a given chunk is a group.
semanticHead is an element of a syntactic group, denoted as a semantic head.

A few of design decisions reflected in these procedures require further comment. First of all, they differ a lot from the studies that involve a definition represented with a bag of words (Dakka & Cucerzan, 2008; Ruiz-Casado et al., 2005; Balasuriya et al., 2009). Here, a certain definition structure is assumed, i.e. a series of nominal groups divided by separators. What is more, as the full stop belongs to them, the series may continue beyond a single sentence, which has improved recall in preliminary experiments. Availability of a shallow parsing layer and group lemmatisation allows to query WordNet by syntactic groups instead of single nouns, as in work of Toral & Muñoz (2006). As word order is relatively free in Polish, a nominal group cannot be assumed to end with a noun, like Kazama & Torisawa (2007) did. Instead, a semantic head of a group is used.

Finally, the problem of lack of word sense disambiguation remains – the line *getLexemes(lemma).synset(0)* means that always a synset connected to the first meaning of a lexeme is selected. We assume that it corresponds to the most common meaning, but that is not always the case – in our example at figure 3 <prezydent.1, prezydent miasta.1> (president of a city, i.e. mayor) precedes <prezydent.2> (president of a country, the obvious meaning). However, it does not have to harm QA performance as far as the question analysis module (section 2.4) functions analogously, e.g. in case of a question beginning with *który prezydent. . .* (which president. . .). Therefore, the decision has been motivated by relatively good performance of this solution in previously performed experiments on question analysis (Przybyła, 2013a). It also works in other applications, e.g. gazetteers generation (Toral & Muñoz, 2006).

To assess quality of the entity library, its content has been compared with synsets manually extracted from randomly selected 100 Wikipedia articles. 95 of them contain a description of an

entity in the first paragraph. Among those, DeepER entity library includes 88 (per-entity recall 92.63 per cent). 135 synsets have been manually assigned to those entities, while the corresponding set in library contains 133 items. 106 of them are equal (per-synset precision 79,70 per cent), while 13 differ only by word sense. 16 of manually extracted synsets hove no counterpart in the entity library (per-synset recall 88.15 per cent), which instead includes 14 false synsets.

3.3 Entity Recognition

An entity recognition step is performed within the question answering process and aims at selecting all entity mentions in a given annotated document. Before it begins, the entity library is read into a PATRICIA trie, a very efficient prefix tree. In this structure, every entity name becomes a key for storing a corresponding list of entities.

When a document is ready for analysis, it is searched for strings that match any of the keys in the trie. The candidate chunks (sequences of segments) come from three sources:

1. lemmata of words and syntactic groups,
2. sequences of words in surface forms (as they appear in text),
3. sequences of words in base forms (lemmata).

The last two techniques are necessary, because a nominal group lemmatisation often fails, especially in case of proper names. Their rich inflection in Polish (Przepiórkowski, 2007) means that a nominal suffix of an entity may be hard to predict. Therefore, a chunk is considered to match an entity name if:

– they share a common prefix,
– an unmatched suffix in neither of them is longer than 3 characters,
– the common prefix is longer than the unmatched chunk suffix.

Given a list of entity mentions, RAFAEL checks their compatibility with a question model. Two of its constituents are taken into account: a general question type and a synset. An entity mention agrees with NAMED_ENTITY type if its first segment starts with a capital letter and always agrees with UNNAMED_ENTITY. To pass a semantic agreement test, the synset of the question model needs to be a (direct or indirect) hypernym of one of the synsets assigned to the entity. For example, list of synsets assigned to entity *Jan III Sobieski* contains <król.1> (king), so it matches a question focus <władca.1, panujący.1, hierarcha.2, pan.1> (ruler) through a hypernymy path <władca.1, panujący.1, hierarcha.2, pan.1> → <monarcha.1, koronowana głowa.1> (monarch) → <król.1>. All the mentions of entities satisfying these conditions are returned for further processing.

4 Evaluation

Evaluation of RAFAEL is typical for factoid QA systems: given a knowledge base and and questions, its responses are compared to the expected ones, prepared in advance. Section 4.1 describes data used in this procedure, whereas section 4.2 explains how an automatic evaluation is possible without human labour.

4.1 Data

The Polish Wikipedia serves as a knowledge base. It has been downloaded from a project site as a single database dump at 03.03.2013, from which plain text files have been extracted using *Wikipedia Extractor* 2.2 script[5]. It means that only plain text is taken into account – without lists, infoboxes, tables, etc. This procedure leads to a corpus with 895,486 documents, containing 168,982,550 segments, which undergo the annotation process, described in section 2.3.

The questions that are to be answered with the knowledge base come from two separate sets:

---

[5] `http://medialab.di.unipi.it/wiki/Wikipedia_Extractor`

1. *Development set* bases on 1500 (1130 after filtering[6]) questions from a Polish quiz TV show, called *Jeden z dziesięciu* (Karzewski, 1997). It was involved in previous experiments (Przybyła, 2013b,a).
2. *Evaluation set* bases on an open dataset for Polish QA systems, published by Marcińczuk et al. (2013). It has been gathered from *Did you know...* column, appearing in the main page of the Polish Wikipedia. It contains 4721 questions, from which 1000 have been analysed, which resulted in 576 satisfying the task constrains, given in chapter 2.

Table 1 shows a distribution of different question types and named entity types in the sets.

General type	Named entity type	Development	Final evaluation
WHICH	-	2.39%	0.17%
TRUEORFALSE	-	2.21%	0.87%
MULTIPLE	-	2.57%	8.33%
OTHER NAME	-	2.04%	0.87%
UNNAMED_ENTITY	-	32.30%	16.67%
	PLACE	2.83%	9.03%
	CONTINENT	0.35%	0.17%
	RIVER	0.97%	0.17%
	LAKE	0.80%	0.00%
	MOUNTAIN	0.35%	0.52%
	RANGE	0.18%	0.17%
	ISLAND	0.44%	0.17%
	ARCHIPELAGO	0.18%	0.00%
	SEA	0.18%	0.00%
	CELESTIAL_BODY	0.71%	0.17%
	COUNTRY	4.60%	0.35%
	STATE	0.62%	0.17%
	CITY	4.69%	0.35%
	NATIONALITY	1.06%	0.17%
	PERSON	22.92%	36.81%
NAMED_ENTITY	NAME	0.97%	0.00%
	SURNAME	0.88%	0.00%
	BAND	0.53%	0.17%
	DYNASTY	0.53%	0.17%
	ORGANISATION	1.86%	3.47%
	COMPANY	0.18%	0.52%
	EVENT	0.97%	1.39%
	TIME	0.18%	3.82%
	CENTURY	0.80%	0.00%
	YEAR	3.01%	0.69%
	PERIOD	0.09%	0.17%
	COUNT	2.74%	7.64%
	QUANTITY	0.53%	1.74%
	VEHICLE	0.88%	1.74%
	ANIMAL	0.09%	0.00%
	TITLE	3.36%	3.47%

Table 1. A distribution of different general types and named entity types in development (1130 questions) and final evaluation (576 questions) sets.

To each of the questions from both sets some information has been assigned manually. It includes an identification number, an expected answer string, a general question type, a named entity type (if applicable) and an expected source document. Table 2 contains several exemplary questions from the development set.

---

[6] The questions that were filtered out either do not belong to *factoid* category, as defined at the beginning of chapter 2 (e.g. demand calculations, translations or long explanations) or lack answer in the Polish Wikipedia.

| | Question | |
Question type	Source article	Answer
	W którym roku umarł Stefan Żeromski?	
	What year did Stefan Żeromski die?	
NAMED_ENTITY :YEAR	Stefan Żeromski	1925
	Jakie organella nadają barwę korzeniom marchwi?	
	What organelles does the carrot take its colour from?	
UNNAMED_ENTITY	Chromoplast	Chromoplasty *Chromoplasts*
	Jakiego wyznania jest większość mieszkańców Liechtensteinu?	
	What is the major confession in Liechtenstein?	
UNNAMED_ENTITY	Liechtenstein	Katolicyzm *Catholicism*
	Który z filozofów był twórcą „atomizmu"?	
	Which philosopher formulated the atomic theory?	
NAMED_ENTITY :PERSON	Demokryt	Demokryt z Abdery *Democritus of Abdera*
	Czy Jacques Brel pochodził z Francji?	
	Was Jacques Brel born in France?	
TRUEORFALSE	Jacques Brel	Nie *No*

Table 2. Exemplary questions with their types (general and named entity), expected source articles and answers.

The additional information (question types and expected documents) makes it possible to evaluate only selected modules of the whole QA system. For example, we could test question classification by comparing results against given question types or entity selection by analysing only the relevant document.

4.2 Automatic Evaluation

Thanks to availability of the DeepER entity library, it is possible to automatically perform answer evaluation for all the question types that are recognised by this technique (UNNAMED_ENTITY and NAMED_ENTITY excluding dates, numbers and quantities).

Both an expected and obtained answer are represented as short strings, e.g. *Bronisław Komorowski*. However, it does not suffice to check their exact equality. That is caused by existence of different names for one entity (*Bronisław Maria Komorowski* or *Komorowski*), but also rich nominal inflection (*Komorowskiego, Komorowskiemu, . . .*).

In fact, we want to compare *entities*, not *names*. Hence, deep entity recognition is a natural solution here. To check correctness of an answer, we use it as an input for the recognition process, described in section 3.3. Then, it is enough to check whether the expected answer appears in any of lists of names, assigned to the recognized entities. For example, let us consider a question: *Kto jest obecnie prezydentem Polski?* (Who is the current president of Poland?) with expected answer *Bronisław Komorowski* and a system answer *Komorowski*. The DeepER process finds many entities in the string (all the persons bearing this popular surname). One of them is the question goal, hence, has *Bronisław Komorowski* in its list of names.

As the process of entity recognition is imperfect, so is the automatic evaluation. However, it still lets us to notice general trends in answering performance with respect to several factors. Of course, the final evaluation needs to be checked manually.

5 Results

As mentioned in previous section, the results consist of two groups: experiments, showing an influence of some aspects of algorithm on performance, and a final assessment. Both use the Polish Wikipedia as a knowledge base, whereas the questions asked belong to development and evaluation

sets, respectively. In this section, *recall* measures percentage of questions, to which RAFAEL gave any answer, whereas *precision* denotes percentage of question answered correctly.

When analysing results of different entity recognition techniques, we need to remember that they strongly rely on output of the question analysis, which is not perfect. In particular, tests show that 15.65 per cent of questions is assigned to wrong type and 17.81 per cent search results do not include the expected document (Przybyła, 2013a). The entity recognition (ER) stage, a focus of this work, is very unlikely to deliver valid answers in these cases. However, as the expected question type and source document are available in question metadata, it is possible to *correct* results of question analysis by artificially replacing a wrong type and/or adding the expected document to the retrieved set. In that way the ER modules could be evaluated, as if question analysis worked perfectly. Note that this approach slightly favours NER-based solutions as the question metadata contains general types and named entity types but lack focus synsets, used by DeepER.

5.1 Experiments

The goal of the first experiment is to test how number a of documents retrieved from the search engine and analysed by the entity recognition techniques, influences the performance. Question classification errors have been bypassed as described in the previous paragraph. Additionally, two versions have been evaluated: with and without corrections of a retrieved set of documents. Figure 4 demonstrates results for different entity recognition techniques.

As we can see, if a retrieved set contains the desired article, adding new documents slightly increases recall, while precision drops observably. That is because additional irrelevant documents usually introduce noise. However, in some cases they are useful, as increasing recall indicates. On the other hand, if we have no guarantee of presence of the expected document in a list, it seems more desirable to extend it, especially for small sizes. For sets bigger than 50 elements, the noise factor again dominates our results. Judging by F1 measure, the optimal value is 20 documents.

When it comes to the comparison, it should be noted that DeepER performs noticeably better than traditional NER. The gain in precision is small, but recall is almost twice as big. It could be easily explained by the fact that the NER solutions are unable to handle UNNAMED_ENTITY type, which accounts for 36 per cent of the entity questions.

It is also worthwhile to check how the system performs while using different values of minimal confidence rate (Jaccard similarity), as described in section 2.7. It could become useful when we demand higher precision and approve lower recall ratio. The plot in figure 5 shows answering performance using DeepER with corrected question analysis with respect to the minimal confidence rate. Generally, the system behaves as expected, but the exact values disappoint. The precision remain at a level of 25-40 per cent up to confidence 0.75, where in turn recall drops to 0.35 per cent only. Values of F1 measure suggest that 0.2 is the highest sensible confidence rate.

One more parameter worth testing, explained in section 2.7, is the context generation strategy. To find the entity with a context most similar to a question content, we could analyse a single sentence, where it appears, or a sequence of words of a predefined length[7]. For both of these solutions, we could also add a document title, as it is likely to be referred to by anaphoric expressions. Figure 6 shows the value of precision (recall does not depend on context) for these four solutions.

We can see that inclusion of a title in a context helps to achieve a better precision. The impact of anaphoric reference to title emerges clearly in case of flexible context – the difference grows with context size. Quite surprisingly, for the optimal context length (1.5 * question size), it is on the contrary. However, because of the small difference between the techniques including title, for the sake of simplicity, the single sentence is used in the final evaluation.

5.2 Final System Evaluation

To impose a realistic challenge to the system, the evaluation set, used at this stage, substantially differs from the one used during the development (see section 4.1). A configuration for the final evaluation has been prepared based on results of the experiments. All of the tested versions share the following features:

– no question analysis corrections,

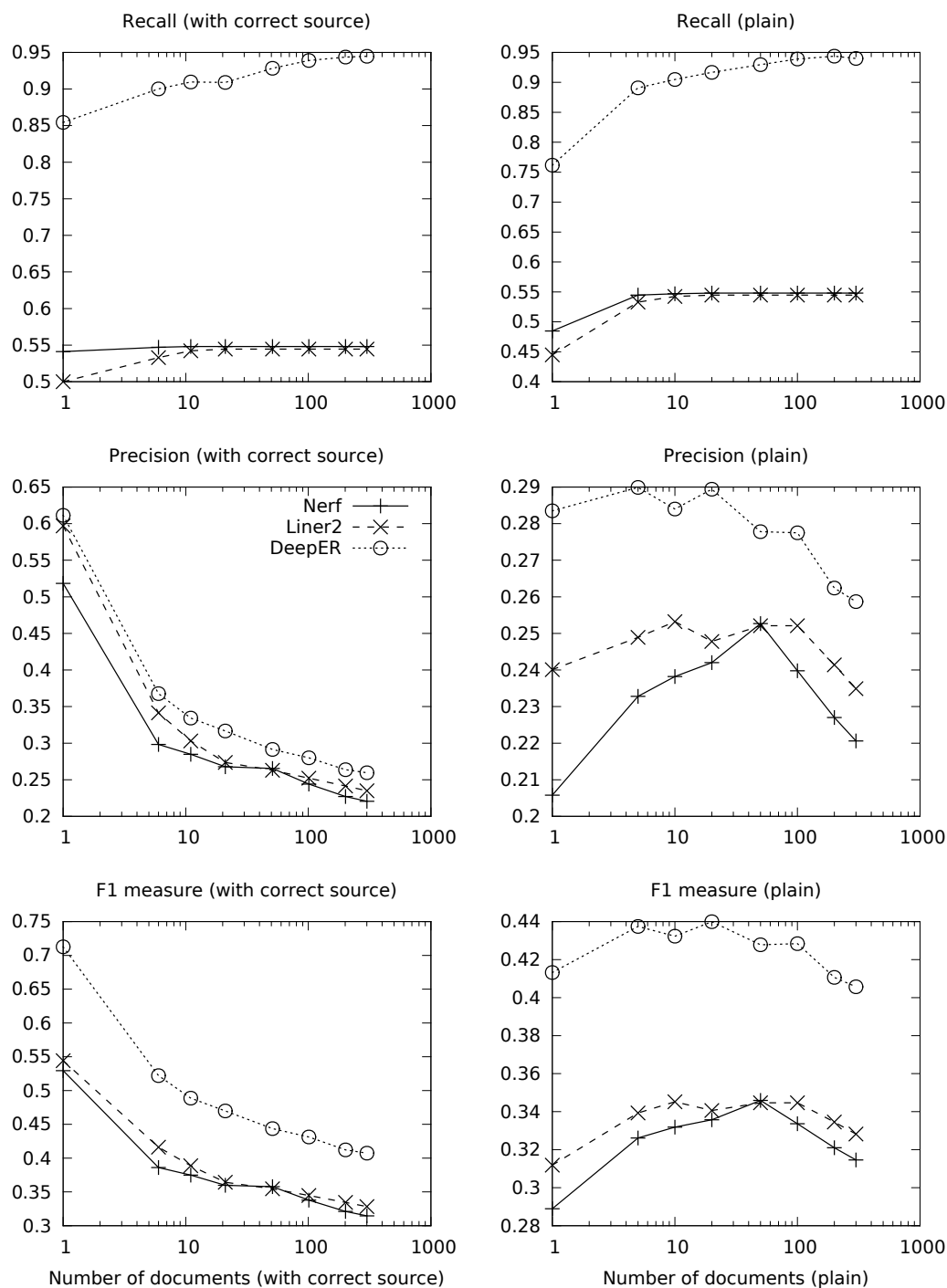[7] Expressed as a multiplication of question content size.

Fig. 4. Question answering performance with respect to size of a retrieved set of documents, undergoing a full analysis. Two versions are considered – with and without guaranteed presence of an article, containing the desired information, in a set. The results for different entity recognition techniques– traditional NER (*Nerf, Liner2*) and *DeepER*.
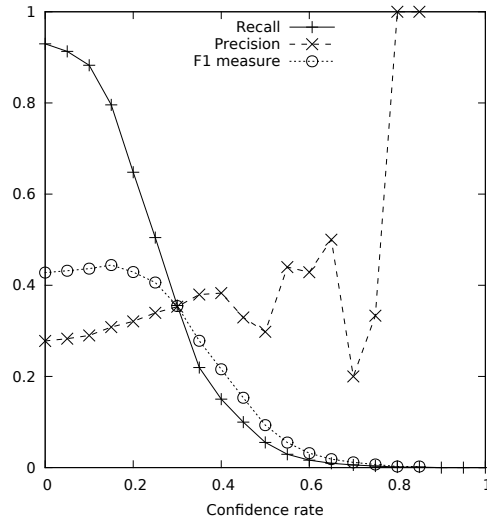
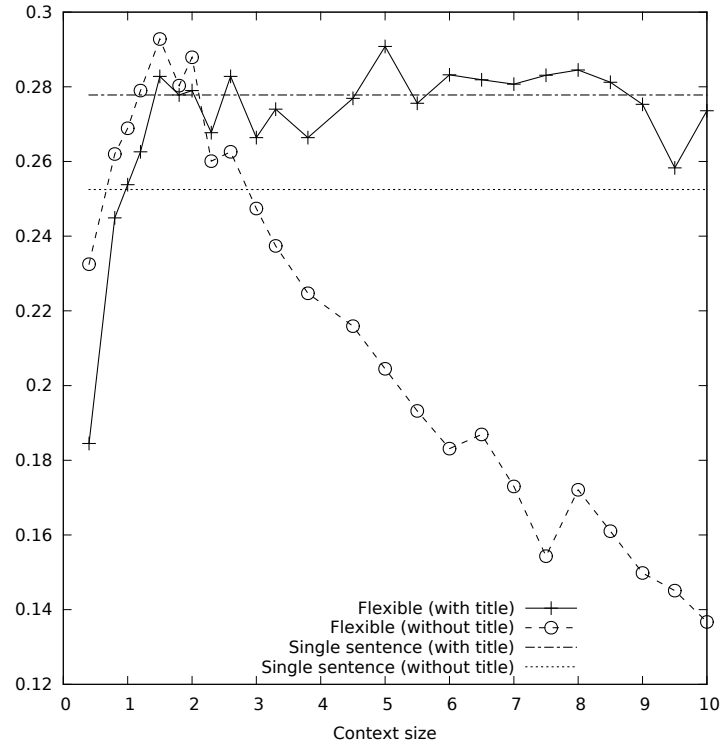**Fig. 5.** RAFAEL performance with respect to minimal confidence rate. Results computed using DeepER with corrected question type and corrected list of 50 documents.

Fig. 6. Question answering performance for different context generation strategies: single sentence and sequence of segments of certain length. Both types considered with and without an article title added.

- question classification and query generation solutions which proved best in the previous experiments (see section 2.4),
- a retrieved set of documents including 20 articles,
- no minimal confidence,
- singe sentence context with title.

Tested solutions differ with respect to entity recognition only; RAFEL variants based on the following options are considered:

- quantities recognizer (*Quant*),
- traditional NER solutions: *Nerf* and *Liner2*,
- deep entity recognition (*DeepER*),
- hybrid approach, where entity mentions were gathered from all the above sources.

ER solution	Recall value	σ	Precision manual value	σ	auto value	F1 measure value	σ	MRR value	σ
Quant	9.55%	1.19%	27.27%	5.6%	-	0.1415	0.0162	26.71%	4.86%
Nerf	56.25%	2.12%	34.88%	2.73%	-	0.4306	0.0213	33.66%	2.29%
Liner2	45.31%	2.05%	39.08%	2.90%	34.10%	0.4197	0.0188	41.36%	2.70%
DeepER	72.92%	1.88%	35.24%	2.23%	33.89%	0.4751	0.0214	32.80%	1.99%
Hybrid	89.58%	1.24%	33.14%	2.01%	-	0.4838	0.0221	35.57%	1.88%

Table 3. Question answering accuracy of RAFEL with different entity recognition strategies: quantities only (*Quant*), traditional NER (*Nerf, Liner2*), deep entity recognition (*DeepER*) and their combination (*Hybrid*).

Table 3 shows results of the final evaluation, expressed by recall, precision, F1 measure and Mean Reciprocal Rank (MRR)[8]. Standard deviations of these values have been obtained by bootstrap resampling of the test set. Additionally, precision obtained by automatic evaluation has been added, where applicable[9]. As we can see, only a small percentage of questions is handled by the quantitative entities recognition. NER-based solutions deal with slightly more (*Nerf*) or less (*Liner2*) than a half of the questions. When using DeepER, the recall ratio rises to 73 per cent while the precision does not differ significantly. That is because UNNAMED_ENTITY questions (unreachable for traditional NER) account for a substantial part of the test set. The maximum recall is obtained by the hybrid solution (90 per cent) but it comes at a cost of lower precision (33 per cent). On the other hand, when we take the whole ranking lists into account, traditional NERs seem to perform better (in terms of MRR).

As expected, the automatic evaluation underestimates precision, but the difference remains below 5 per cent. Judging by F1 measure, the hybrid solution seems to beat the others.

6 Discussion

The main strength of DeepER compared to NER, according to results shown in figure 3, is much higher recall. Table 4 shows examples of questions, to which only DeepER provides a correct answer. As we can see (notice question foci in the table), they could not be assigned to any of the traditional NE categories.

The other striking fact in the results is low precision. A part of the wrong answers was inspected and most of the errors seem to result from the following phenomena:

- Information given in question in concise form, in article appears scattered over many sentences.

---

[8] Evaluated automatically due to length of ranking lists.
[9] Results obtained using *Quant* and *Nerf* may be expressed by quantities and dates, which are not supported by the entity-based automatic evaluation

	Question	
Question focus		**Answer**
	Który żleb uważany jest za najbardziej lawiniasty w całych Karpatach?	
	Which gulley is considered the most avalanche-prone in the Carpathians?	
żleb		Pusty Żleb
gulley		
	Jaki owad pożera liście owadożernej rosiczki?	
	What insect feeds on leaves of carnivorous sundew?	
owad		Piórolotek bagniczek
insect		*Buckleria paludum*
	Jaki przyrząd pozwala na pomiar współczynnika załamania światła?	
	What apparatus is used to measure the refractive index?	
przyrząd		Refraktometr Abbego
apparatus		*Abbe refractometer*
	Jaki związek chemiczny służy do otrzymywania boru o wysokiej czystości?	
	What chemical compound is used to obtain boron of high purity?	
związek chemiczny		Jodek boru
chemical compound		*Boron triiodide*
	Jaką metodą łamano szyfry Enigmy przed wynalezieniem cyklometru?	
	What method was used to decrypt Enigma cipher before the advent of the cyclometer?	
metoda		Metoda rusztu
method		*Grill*

Table 4. Examples of questions which have been handled and answered correctly only with the DeepER approach. Their foci lie beyond areas covered by the NE categories.

- Some of the pairs of words in question content and answer context are very related morphologically or semantically, but differ in lemmata, e.g. *łowca* (huntsman) and *łowiectwo* (huntsmanship) or *myśliwy* (huntsman).
- Because of rich inflection in Polish, a tagger quite often fails to select the correct lemma of a word, especially in case of proper names.
- If several entities of a desired type appear in one sentence, the bag-of-words model does not suffice to select the best one.

The entity recognizers also introduce errors typical for them:

- *Quant* ignores quantity types; therefore, a year is frequently returned as an answer to questions beginning with *How many . . . .*
- *Nerf* and *Liner2* have insufficient recall – they recognize only a fraction of entities of a desired type.
- *DeepER* suffers from lack of word sense disambiguation, which impedes WordNet-based inference.

The last remark applies also to other techniques. For example, consider a word *kot*, which means *a cat*. However, it is also a name of a journal, a lake, a village, a badge (*KOT*), a surname of 10 persons in the Polish Wikipedia and much more. A human would usually assume the most common meaning (a cat), but the system treats them as equally probable. It introduces noise in the process, as such an entity matches many types of questions.

Another thing that demands explanation is a difference in precision of answers found using *Liner2* and DeepER: in evaluation set the latter does not maintain its advantage from development set. It could be explained by different compositions of the question sets (table 1) – the development one contains much more questions beginning with ambiguous pronouns, followed by a question focus, e.g. *Który poeta. . .* (which poet), thus providing a precise synset (a poet) for deep entity recognition. Members of the evaluation set much more frequently begin with pronouns like *Kto . . .* (who), where a synset corresponds to a general NE type (a person).

As RAFAEL is the first Polish QA system, able to answer by entities instead of documents, we can not compare it directly to any other solution. However, the evaluation set has been created based on questions published by Marcińczuk et al. (2013) and used for evaluation of a document retrieval system (Marcińczuk et al., 2013). Their baseline configuration achieved a@1 (percentage of

questions answered by the first document, corresponds to precision in table 3) equal 26.09 per cent. By taking into account proximity of keyword matches (MCSW method), they improved the result to 38.63 per cent. We can see that RAFAEL, despite solving much more challenging problem, in all configurations obtains better precision than baseline; using *Liner2* it beats even the best method tested on this set (MCSW).

The results suggest two possible directions of future work to improve performance of RAFAEL. Firstly, involving semantics in sentence matching could solve some of the problems mentioned above. There are a lot of techniques in that area, also in QA systems (see a variety of them used by Yih et al. (2013)), but their implementation in a morphologically rich language would require a thorough study. For example, there exist techniques computing a semantic similarity based on a WordNet graph (Moldovan & Novischi, 2002), which is available for Polish and proved very useful in this study. Secondly, the relatively good performance of hybrid ER indicates that it may be good to apply different entity recognizer to different questions. For example, we could evaluate them for each question type separately and select the one that performs best for a given one. However, it would require much more training data to have a substantial number of questions of each type, including the scarce ones (observe sparsity of table 1).

When it comes to DeepER, word ambiguity seem to be the main issue for future efforts. Of course, a full-lexicon precise word-sense disambiguation tool would solve the problem, but we can't expect it in near future. Instead, we could select a synset somewhere in a path between a focus synset and a named entity type. In the example from figure 3 rather than choosing between <prezydent.1, prezydent miasta.1> (president of a city) and <prezydent.2> (president of a country) we could use <urzędnik.1, biuralista.1> (official), which covers both meanings.

7   Conclusions

This paper introduces RAFAEL, a complete open-domain question answering system for Polish. It is capable of analysing a given question, scanning a large corpus and extracting an answer, represented as a short string of text.

In its design, the focus has been on entity recognition techniques, used to extract all the entities compatible with a question from a given text. Apart from the traditional named entity recognition, differentiating between several broad categories of NEs, a novel technique, called Deep Entity Recognition (DeepER), has been proposed and implemented. It is able to find entities belonging to a given WordNet synset, using an entity library, gathered by interpreting definitions from encyclopaedia.

Automatic evaluation, provided by DeepER approach, has let to perform several experiments, showing answering accuracy with respect to different parameters. Their conclusions have been used to prepare final evaluation, which results have been checked manually. They suggest that the DeepER-based solution yields similar precision to NER, but is able to answer much more questions, including those beyond the traditional categories of named entities.

Appendix A: Named Entity Recognition in RAFAEL

As mentioned in section 2.6, apart from DeepER, RAFAEL employs also traditional NER-based solutions for entity recognition: *NERF*, *Liner2* and *Quant*. Each of them uses its own typology of named entities, which covers only a part of the types, enumerated in section 2.4. Table 5 shows a correspondence between these types. As we can see, there are a few problems:

1. Many of NE types are not covered by neither *NERF* nor *Liner2*,
2. For all geographical types, *NERF* has only one type *geogName*, which may affect QA precision,
3. In case of *CENTURY* and *YEAR*, *NERF* recognizes only a more general type *date*, from which they may be inferred,
4. *Liner2* does not differentiate between *NAME* and *SURNAME*, classifying both as parts of *person_nam*.

The problems 3 and 4 are solved by an additional postprocessing code, extracting *CENTURY* from *date* and *NAME* and *SURNAME* from *person_nam* entities. In case of multi-segment person entities it assumes that the first and last word correspond to first and last name, respectively.

Question NE type	NERF type	Liner2 type	Quant type
PLACE	placeName:*		
CONTINENT		continent_nam	
RIVER		river_nam	
LAKE			
MOUNTAIN			
RANGE	geogName	mountain_nam	
ISLAND		island_nam	
ARCHIPELAGO			
SEA		sea_nam	
CELESTIAL_BODY		astronomical_nam	
COUNTRY	placeName:country	country_nam	
STATE	placeName:region	admin1_nam	
		admin2_nam	
		admin3_nam	
		historical_region_nam	
		country_region_nam	
CITY	placeName:settlement	city_nam	
NATIONALITY	placeName:country	nation_nam	
PERSON	persName		
NAME	persName:forename	person_nam	
SURNAME	persName:surname		
BAND	orgName	band_nam	
DYNASTY	persName:addName		
ORGANISATION	orgName	organization_nam	
		institution_nam	
		political_party_nam	
COMPANY	orgName	company_nam	
EVENT		event_nam	
TIME			
CENTURY	date		
YEAR			
PERIOD			quantity
COUNT			number
QUANTITY			quantity
VEHICLE			
ANIMAL			
TITLE		title_nam	
		media_nam	

Table 5. Correspondence between named entity types from question analysis and supported by different NER solutions.

While *NERF* and *Liner2* are standalone NER tools and details of their design are available in previously mentioned publications, *Quant* has been created specifically for RAFAEL. To find numbers, it annotates all chains of segments according to a predefined pattern, which accepts the following types of segments:

1. **(0-9)+** – a string consisting of digits only,
2. **.** – a period; a digit group separator in Polish ,
3. **,** – a comma; a decimal mark in Polish ,
4. **num** – a verbal expression of number, i.e. segments tagged as numerals.

The pattern is matched in greedy mode, i.e. it adds as many new segments as possible. It could recognise expressions like *10 tysięcy* (10 thousand), *kilka milionów* (several million), *10 000* or *1.698,88* (1,698.88).

Quantity is a sequence of segments, recognised as a number, followed by a unit of measurement. To check whether a word denotes a unit of measurement, the *plWordNet* is searched for lexemes equal to its base. Then it suffices to check whether it belongs to a synset, having <jednostka miary 1> (unit of measurement) as one of (direct or indirect) hypernyms, e.g. *piętnaście kilogramów* (fifteen kilograms) or *5 000 watów* (5 000 watts).

Acknowledgments

Bibliography

Acedański, S. (2010). A morphosyntactic Brill Tagger for inflectional languages. In *Proceedings of the 7th international conference on Advances in Natural Language Processing (IceTAL'10)*, (pp. 3–14). Springer-Verlag.

Ahn, D., Jijkoun, V., Mishne, G., Müller, K., De Rijke, M., & Schlobach, S. (2004). Using Wikipedia at the TREC QA Track. In Voorhees, E. M. & Buckland, L. P. (Eds.), *Proceedings of The Thirteenth Text REtrieval Conference (TREC 2004)*.

Armenska, J., Tomovski, A., Zdravkova, K., & Pehcevski, J. (2010). Information Retrieval Using a Macedonian Test Collection for Question Answering. In *Proceedings of the 2nd International Conference ICT Innovations*, (pp. 205–214). Springer-Verlag.

Balasuriya, D., Ringland, N., Nothman, J., Murphy, T., & Curran, J. R. (2009). Named entity recognition in Wikipedia. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, (pp. 10–18)., Association for Computational Linguistics.

Brill, E., Dumais, S., & Banko, M. (2002). An analysis of the AskMSR question-answering system. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02*, volume 10, (pp. 257–264). Association for Computational Linguistics.

Ciaramita, M. & Altun, Y. (2006). Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing - EMNLP '06*, (pp. 594). Association for Computational Linguistics.

Dakka, W. & Cucerzan, S. (2008). Augmenting Wikipedia with Named Entity Tags. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*.

Dang, H. T., Kelly, D., & Lin, J. (2007). Overview of the TREC 2007 Question Answering track. In *Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007*.

Degórski, L. (2012). Towards the Lemmatisation of Polish Nominal Syntactic Groups Using a Shallow Grammar. In Bouvry, P., Kłopotek, M. A., Leprévost, F., Marciniak, M., Mykowiecka, A., & Rybiński, H. (Eds.), *Proceedings of the International Joint Conference on Security and Intelligent Information Systems*, volume 7053 of *Lecture Notes in Computer Science*, (pp. 370–378). Springer-Verlag.

Duclaye, F., Sitko, J., Filoche, P., & Collin, O. (2002). A Polish Question-Answering System for Business Information. In *BIS 2002, 5th International Conference on Business Information Systems, Poznań, Poland, 24-25 April 2002*, (pp. 209–212).

Ferrucci, D. A., Brown, E., Chu-carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., Lally, A., Murdock, J. W., Nyberg, E., Prager, J., Schlaefer, N., & Welty, C. (2010). Building Watson: An Overview of the DeepQA Project. *AI Magazine*, *31*(3), 59–79.

Galambos, L. (2001). Lemmatizer for Document Information Retrieval Systems in JAVA. In *Proceedings of the 28th Conference on Current Trends in Theory and Practice of Informatics (SOFSEM 2001)*, (pp. 243–252). Springer-Verlag.

Harabagiu, S., Moldovan, D., Pasca, M., Mihalcea, R., Surdeanu, M., Bunescu, R., Gîrju, R., Rus, V., & Morarescu, P. (2001). The role of lexico-semantic feedback in open-domain textual question-answering. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL '01*, (pp. 282–289). Association for Computational Linguistics.

Hovy, E., Gerber, L., Hermjakob, U., Junk, M., & Lin, C.-Y. (2000). Question Answering in Webclopedia. In *Proceedings of The Ninth Text REtrieval Conference (TREC 2000)*.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, *37*, 547–579.

Karzewski, M. (1997). *Jeden z dziesięciu - pytania i odpowiedzi*. Muza SA.

Katz, B., Lin, J., Loreto, D., Hildebrandt, W., Bilotti, M., Felshin, S., Fernandes, A., Marton, G., & Mora, F. (2003). Integrating Web-based and corpus-based techniques for question answering. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*.

Kazama, J. & Torisawa, K. (2007). Exploiting Wikipedia as External Knowledge for Named Entity Recognition. In *In Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (2007)*, (pp. 698–707). Association for Computational Linguistics.

Konopík, M. & Rohlík, O. (2010). Question Answering for Not Yet Semantic Web. In *Proceedings of the 13th International Conference on Text, Speech and Dialogue (TSD 2010)*, (pp. 125–132). Springer-Verlag.

Kopeć, M. & Ogrodniczuk, M. (2012). Creating a Coreference Resolution System for Polish. In *The eighth international conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA).

Lee, C., Wang, J.-H., Kim, H.-J., & Jang, M.-G. (2005). Extracting Template for Knowledge-based Question-Answering Using Conditional Random Fields. In *Proceedings of the 28th Annual International ACM SIGIR Workshop on MFIR*, (pp. 428–434). ACM.

Li, X. & Roth, D. (2002). Learning Question Classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*, volume 1 of *COLING '02*. Association for Computational Linguistics.

Lombarović, T., Šnajder, J., & Bašić, B. D. (2011). Question Classification for a Croatian QA System. In *Proceedings of the 14th International Conference on Text, Speech and Dialogue (TSD 2011)*, (pp. 403–410). Springer-Verlag.

Mann, G. S. (2002). Fine-grained proper noun ontologies for question answering. In *Proceedings of the 2002 workshop on Building and using semantic networks (SEMANET '02)*, volume 11, (pp. 1–7). Association for Computational Linguistics.

Marcińczuk, M. & Janicki, M. (2012). Optimizing CRF-based Model for Proper Name Recognition in Polish Texts. In *Proceedings of CICLing 2012, Part I*, (pp. 258–269). Springer-Verlag.

Marcińczuk, M., Ptak, M., Radziszewski, A., & Piasecki, M. (2013). Open dataset for development of Polish Question Answering systems. In *Proceedings of the 6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. Wydawnictwo Poznańskie, Fundacja Uniwersytetu im. Adama Mickiewicza.

Marcińczuk, M., Radziszewski, A., Piasecki, M., Piasecki, D., & Ptak, M. (2013). Evaluation of a Baseline Information Retrieval for a Polish Open-domain Question Answering System. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2013)*, (pp. 428–435). Association for Computational Linguistics.

Maziarz, M., Piasecki, M., & Szpakowicz, S. (2012). Approaching plWordNet 2.0. In *Proceedings of the 6th Global Wordnet Conference*.

Moldovan, D., Harabagiu, S., Paşca, M., Mihalcea, R., Gîrju, R., Goodrum, R., & Rus, V. (2000). The structure and performance of an open-domain question answering system. In *Proceedings of*

the 38th Annual Meeting on Association for Computational Linguistics - ACL '00, (pp. 563–570). Association for Computational Linguistics.

Moldovan, D. & Novischi, A. (2002). Lexical chains for question answering. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*. Association for Computational Linguistics.

Oh, J.-h., Torisawa, K., Hashimoto, C., Sano, M., Saeger, S. D., & Ohtake, K. (2013). Why-Question Answering using Intra- and Inter-Sentential Causal Relations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, (pp. 1733–1743). Association for Computational Linguistics.

Osenova, P., Simov, A., Simov, K., Tanev, H., & Kouylekov, M. (2004). Bulgarian-english question answering: adaptation of language resources. In Peters, C., Clough, P., Gonzalo, J., Jones, G. J. F., Kluck, M., & Magnini, B. (Eds.), *Proceedings of the 5th conference on Cross-Language Evaluation Forum: multilingual Information Access for Text, Speech and Images (CLEF'04)*, volume 3491 of *Lecture Notes in Computer Science*, (pp. 458–469). Springer-Verlag.

Peshterliev, S. & Koychev, I. (2011). Semantic Retrieval Approach to Factoid Question Answering for Bulgarian. In *Proceedings of the 3rd International Conference on Software, Services and Semantic Technologies (S3T 2011)*, (pp. 25–32). Springer-Verlag.

Piechociński, D. & Mykowiecka, A. (2005). Question answering in Polish using shallow parsing. In Garabík, R. (Ed.), *Computer Treatment of Slavic and East European Languages: Proceedings of the Third International Seminar, Bratislava, Slovakia*, (pp. 167–173). VEDA: Vydava- tel'stvo Slovenskej akadéme vied.

Ponzetto, S. P. & Strube, M. (2007). Deriving a Large Scale Taxonomy from Wikipedia. *Artificial Intelligence*, *22*, 1440–1445.

Przepiórkowski, A. (2007). Slavonic information extraction and partial parsing. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing Information Extraction and Enabling Technologies - ACL '07*. Association for Computational Linguistics.

Przepiórkowski, A. (2008). *Powierzchniowe przetwarzanie języka polskiego*. Warszawa: Akademicka Oficyna Wydawnicza EXIT.

Przepiórkowski, A., Bańko, M., Górski, R. L., & Lewandowska-Tomaszczyk, B. (2012). *Narodowy Korpus Języka Polskiego*. Warszawa: Wydawnictwo Naukowe PWN.

Przybyła, P. (2012). Issues of Polish Question Answering. In Hryniewicz, O., Mielniczuk, J., Penczek, W., & Waniewski, J. (Eds.), *Proceedings of the first conference 'Information Technologies: Research and their Interdisciplinary Applications' (ITRIA 2012)*, (pp. 122–139)., Warsaw. Institute of Computer Science, Polish Academy of Sciences.

Przybyła, P. (2013a). Question Analysis for Polish Question Answering. In *51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop*, (pp. 96–102)., Sofia, Bulgaria. Association for Computational Linguistics.

Przybyła, P. (2013b). Question Classification for Polish Question Answering. In Kłopotek, M. A., Koronacki, J., Marciniak, M., Mykowiecka, A., & Wierzchoń, S. T. (Eds.), *Proceedings of the 20th International Conference on Language Processing and Intelligent Information Systems (LP&IIS 2013)*, (pp. 50–56). Springer-Verlag.

Richman, A. E. & Schone, P. (2008). Mining Wiki Resources for Multilingual Named Entity Recognition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL 2008)*. Association for Computational Linguistics.

Ruiz-Casado, M., Alfonseca, E., & Castells, P. (2005). Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets. *Advances in Web Intelligence*, *3528*, 380–386.

Savary, A., Manicki, L., & Baron, M. (2013). Populating a multilingual ontology of proper names from open sources. *Journal of Language Modelling*, *1*(2), 189–225.

Savary, A. & Waszczuk, J. (2012). Narzędzia do anotacji jednostek nazewniczych. In *Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]* (pp. 225–252). Wydawnictwo Naukowe PWN.

Shapiro, S. C. (1992). *Encyclopedia of Artificial Intelligence*. John Wiley & Sons, Inc.

Simov, K. & Osenova, P. (2005). BulQA: Bulgarian–bulgarian question answering at CLEF 2005. In Peters, C., Gey, F. C., Gonzalo, J., Müller, H., Jones, G. J. F., Kluck, M., Magnini, B., & Rijke, M. (Eds.), *Proceedings of the 6th international conference on Cross-Language Evalution Forum: accessing Multilingual Information Repositories (CLEF'05)*, volume 4022 of *Lecture Notes in Computer Science*, (pp. 517–526). Springer-Verlag.

Solovyev, A. (2013). Dependency-Based Algorithms for Answer Validation Task in Russian Question Answering. In *Proceedings of the 25th International Conference on Language Processing and Knowledge in the Web (GSCL 2013)*, (pp. 199–212). Springer-Verlag.

Tanev, H. (2004). Socrates - a Question Answering prototype for Bulgarian. In N. Nikolov (Ed.), *Recent Advances in Natural Language Processing: Selected Papers from RANLP 2003* (pp. 377–386). John Benjamins.

Toral, A. & Muñoz, R. (2006). A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Toral, A., Muñoz, R., & Monachini, M. (2008). Named Entity WordNet. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008*.

Čeh, I. & Ojsteršek, M. (2009). Developing a question answering system for the slovene language. *WSEAS Transactions on Information Science and Applications*, *6*(9), 1533–1543.

Vetulani, Z. (1988). PROLOG Implementation of an Access in Polish to a Data Base. In *Studia z automatyki, XII* (pp. 5–23). PWN.

Walas, M. (2012). How to answer yes/no spatial questions using qualitative reasoning? In Gelbukh, A. (Ed.), *13th International Conference on Computational Linguistics and Intelligent Text Processing*, (pp. 330–341). Springer-Verlag.

Walas, M. & Jassem, K. (2010). Named entity recognition in a Polish question answering system. In *Intelligent Information Systems*, (pp. 181–191). Publishing House of University of Podlasie.

Walas, M. & Jassem, K. (2011). Spatial reasoning and disambiguation in the process of knowledge acquisition. In *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, (pp. 420–424). Fundacja Uniwersytetu im. Adama Mickiewicza.

Woliński, M. (2006). Morfeusz — a Practical Tool for the Morphological Analysis of Polish. In M. Kłopotek, S. Wierzchoń, & K. Trojanowski (Eds.), *Intelligent Information Processing and Web Mining* (pp. 511–520). Springer-Verlag.

Yih, W.-t., Chang, M.-w., Meek, C., & Pastusiak, A. (2013). Question Answering Using Enhanced Lexical Semantic Models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, (pp. 1744–1753). Association for Computational Linguistics.