Preliminary Evaluation of Passage Retrieval in Biomedical Multilingual Question Answering

Mariana Neves[1], **Konrad Herbst**[1,2], **Matthias Uflacker**[1], **Hasso Plattner**[1]

[1]Hasso-Plattner-Institute at the University of Potsdam, Germany
[2] University of Heidelberg, Germany
marianalaraneves@gmail.com, k.herbst@stud.uni-heidelberg.de

Abstract

Question answering systems can support biologists and physicians when searching for answers in the scientific literature or the Web. Further, multilingual question answering systems provide more possibilities and flexibility to users, by allowing them to write questions and get answers in their native language, and the exploration of resources in other languages by means of machine translation. We present a prototype of a multilingual question answering system for the biomedical domain for English, German and Spanish. Preliminary experiments have been carried out for passage retrieval based on the multilingual parallel datasets of Medline titles released in the CLEF-ER challenge. Two parallel collections of 50 questions for English/German and English/Spanish have been created to support evaluation of the system. Results show that the task is not straightforward, that additional semantic resources are needed to support query expansion and that different strategies might be necessary for distinct languages.

Keywords: question answering, multilingual, biomedicine.

1. Introduction

Question answering systems (QA) are important tools in the biomedical domain to provide answers to questions arisen from scientists and physicians. For instance, biologists frequently look for answers in the scientific literature or in the Web to confirm results obtained in the laboratory, e.g., whether a certain disease could be associated to a particular genetic mutation. However, answers to such questions can be scattered over different scientific publications (abstracts and full text), biological databases and Web pages.

Despite the importance of QA systems for both fields, not much previous work has been carried out for the biomedical domain, when compared to the state-of-art solutions in the medical domain (Athenikos and Han, 2010). Currently, there seems to be only three QA systems available for the biomedical domain (Bauer and Berleant, 2012): EAGLi[1], AskHermes[2] and HONQA[3]. However, both AskHermes and HONQA have a focus on the medical domain and might not be suitable for biologists. Nevertheless, this scenario has been changing in the last couple of years thanks to new initiatives such as the QA4MRE (Morante et al., 2012) and BioASQ (Partalas et al., 2013) challenges which support improvements in biomedical question answering.

Multilingual question answering systems offer a variety of new possibilities to the user, also for the biological domain. Although most of the relevant scientific publications are available in the English language, there is a myriad of other resources in other languages that could be explored, such as non-English biomedical scientific literature (e.g., Scielo[4] for Spanish and Portuguese), as well as the whole Web. Further, a multilingual QA system allows non-native English speakers to pose questions and receive answers in their native language while the system queries English (or any other language) documents by means of machine translation.

Previous research is scarce in the biomedical multilingual question answering field and in natural language processing (NLP) in general. From the QA systems cited above, only HONQA allows questions to be posed in other languages than English, namely Italian and French, whose performance has been evaluated in (Olvera-Lobo and Gutirrez-Artacho, 2011). The biomedical research seems to be always one (or more) step behind state-of-art in NLP for other domains due to the complexity of the matter and the lack of suitable resources for system development and evaluation. For instance, QA systems need to rely in high performing tools for the many pre-processing steps, such as tokenization, part-of-speech tagging and parsing. However, previous research have proved that sometimes specific models might be necessary for some domains, such as biomedicine (McClosky et al., 2010). One important step towards improvements in biomedical NLP is the recent EU-funded Multilingual Annotation of Named Entities and Terminology Resources Acquisition (Mantra) project (cf. Section 2.1.) which has supported improvements in named-entity recognition of biomedical terms during the CLEF-ER challenge last year. Finally, improvements in machine translation for the biomedical have also been recently published (Jimeno Yepes et al., 2013).

We present a prototype of a multilingual question an-

[1] `eagl.unige.ch/EAGLi/`
[2] `http://www.askhermes.org/`
[3] `http://www.hon.ch/QA/`
[4] `http://www.scielo.org/`

swering system for the biomedical domain and perform a preliminary evaluation for English, German and Spanish based on the resources released during the CLEF-ER challenge (Rebholz-Schuhmann et al., 2013). The QA system was developed on the top of SAP HANA, an in-memory database which include built-in text analysis functionalities for eleven languages. Additionally, we have created and made available two parallel collections of 50 questions each for English/German and English/Spanish. As far as we know, there is no multilingual parallel collection of questions nor previous evaluation for multilingual passage retrieval in biomedical QA systems. So far, the most comprehensive passage retrieval evaluation for the biomedical domain has been performed during the TREC 2006 and 2007 Genomics tracks (Hersh and Voorhees, 2009), but whose questions and background collection seems not to be made available after the end of the challenge.

In the next section, we will present an overview of the CLEF-ER resources followed by a description of the multilingual collection of questions. Then, we describe the question answering system which is under development (cf. Section 3.) and present an evaluation over the created collection of questions (cf. Section 4.). Finally, a discussion on the results and future work are presented in Section 5..

2. Data

2.1. CLEF-ER resources

The CLEF-ER challenge took place in 2013 as part of the Mantra project[5] which aims to provide multilingual documents and terminologies for the biomedical domain. For the scope of this challenge, Medline and patent documents have been released in five languages: English, German, French, Spanish and Dutch. Mapping to English documents were provided for all documents in each of the languages other than English but no mapping seems to exist between documents from two other languages, e.g., between Spanish and German.

For the scope of this work, we have utilized the Medline documents collections, which in fact consists only of the title of the documents, as background collection for our QA prototype. We have restricted our work to Spanish and German, given the support of the SAP HANA database for these languages (cf. Section 3.) and the knowledge of the authors on them. Table 1 illustrates examples of corresponding Medline document titles in English, Spanish and German.

Organizers have also released a terminology containing synonyms for terms in the above languages, which has been compiled from three resources: Medical Subject Headings (MeSH), Systematized Nomenclature of Human and Veterinary Medicine (SNOMED-CT) and Medical Dictionary for Regulatory Activities (MedDRA). An example of a term and its corresponding synonyms in other languages is shown in Table 2.

Table 3 summarizes the total number of Medline documents and synonyms per language within the terminology. These values are based on our own measurements after successful database import, thus, there might be discrepancies to the figures provided by the CLEF-ER organizers[6]. The CLEF-ER terminology contains a total of 525,794 terms which can be associated to synonyms in the many of the supported languages. For instance, in Table 2, one concept is shown (C0000119) which contains 7 synonyms: 3 for English, one for French, one for German and two for Spanish.

Resource / Language	English	German	Spanish
Medline documents	1,593,546	719,232	247,655
Synonyms	1,771,498	118,902	622,638

Table 3: Number of Medline documents and synonyms for English, Spanish and German in the CLEF-ER resources.

2.2. Questions collection

The construction of the collection of parallel questions was based on the Medline documents in Spanish and German which contain a corresponding document in English. We chose this approach to allow a comparison between results obtained with German and Spanish with those in English on the same documents.

Batches of 50 documents (titles) were randomly retrieved from the above datasets and titles were manually chosen according to their relevance to the biomedical domain. During the automatic retrieval of candidate documents, we ignored titles which had length lower than 150 characters as they might not contain enough information for question generation. During manual screening of the titles, we avoided documents related only to the medical domain, an area where state-of-art in question answering is more advanced in comparison to the biological one (Athenikos and Han, 2010). In particular for the Spanish questions, we selected titles related to tropical neglected diseases, which is a frequent topic on publications in the Medline collection for this language. Finally, we ignored titles which were not relevant to the biomedical domain, such as “On the effect of religious schools on values of young people.” (document d5582363), or that consisted of reports on a meeting or on the current situation in a particular place, such as “The 5th Annual Meeting of the Swiss Association for Preventive and Restorative Dentistry (SVPR) of 13 November 1999 in Zurich.” (document d10744522).

Questions were manually written in a way that at least one answer could be found in the corresponding document, i.e., the document’s title. However, not all of the information cited in the text was always used and the questions sometimes are more general than the respective text. We have generated only factoid questions, i.e., questions which

[5] https://sites.google.com/site/mantraeu/home

[6] https://sites.google.com/site/mantraeu/access-content

English/German (document d11951797)
Optimal intravascular brachytherapy: safety and radiation protection, reliability and precision guaranteed by guidelines, recommendations and regulatory requirements. Optimale intravaskuläre Brachytherapie. Sicherheit und Strahlenschutz, Zuverlässigkeit und Präzision gewährleistet durch Leitlinien, Empfehlungen und Verordnungen.

English/Spanish (document d18959013)
Impact of the deep breathing maneuver in the gas exchange in the subject with severe obesity and pulmonary arterial hypertension associated to Eisenmenger's syndrome. Impacto de la maniobra de inspiracin profunda en el intercambio gaseoso del sujeto con obesidad severa e hipertensin arterial pulmonar asociada a sndrome de Eisenmenger.

Table 1: Examples of Medline document titles from the CLEF-ER collection for the pairs English/German and English/Spanish.

[Term]
id: C0000119
name: 11-Hydroxycorticosteroids
namespace: mesh_term_from_umls
def: "A group of corticosteroids bearing a hydroxy group at the 11-position." []
synonym: "11 Hydroxycorticosteroids" EXACT SYN_EN [MSH:D015062]
synonym: "11-Hidroxicorticoesteroides" EXACT PREF_ES [MSHSPA:D015062]
synonym: "11-Hidroxicorticosteroides" EXACT SYN_ES [MSHSPA:D015062]
synonym: "11-Hydroxycorticosteroide" EXACT PREF_DE [MSHGER:D015062]
synonym: "11-Hydroxycorticosteroids" EXACT PREF_EN [MSH:D015062, NDFRT:N0000011376]
synonym: "11-Hydroxycorticosteroids [Chemical/Ingredient]" EXACT SYN_EN [NDFRT:N0000011376]
synonym: "11-Hydroxycorticostrodes" EXACT PREF_FR [MSHFRE:D015062]
is_a: C0020343 ! Hydroxycorticosteroids
relationship: has_semantic_type T110 ! Steroid
relationship: has_semantic_type T125 ! Hormone

Table 2: Term "1-Sarcosine-8-Isoleucine Angiotensin II" from the CLEF-ER terminology and its corresponding synonyms in English, French, Spanish and German.

requires one or more specific short answer in return, such as a chemical compound, an organism or a disease. While writing the questions, we tried to rephrase the text, used synonyms for both the named entities and remaining words (whenever possible), changed the word's lexical class (e.g., from verb to noun), and converted passive voice to active voice, or the other way round, following procedures described in (Heilman and Smith, 2010). Synonyms for the lexical terms were supported by making queries to a variety of on-line language-specific dictionaries.

The semantic concepts referred in the text were also, whenever possible, changed to a equivalent synonyms. This task was supported by a variety of on-line resources, such as Wikipedia (for the three languages), NCBI Taxonomy and other web sites which were returned by the Google search engine. Therefore, we did not make use of the thesaurus made available by the CLEF-ER challenge, instead, we have tried to use the resources that the users (biologists) might use while posing questions to a QA system.

Questions were initially written in English and were reviewed by an expert in molecular biotechnology (KH). In a second step, the questions were translated into Spanish and German by the authors, who are either native or have advanced knowledge on the languages. We have also sought to use synonyms for the words and concepts in this step. Finally, we tried to make the questions with similar difficulty level in both languages. Some examples of questions are presented in Table 4 and the list of parallel questions is available for download[7].

English/German (document d6357751)
Which methods can be used to determine the living cell count of cariogenic microorganisms? Welche Methoden bieten sich zur Bestimmung der Lebendzellzahl von kariogenen Mikroorganismen an?

English/Spanish (document d16888692)
What are possible drug targets for eye related infections? Cuáles son los posibles objetivos farmacológicos en infecciones relacionadas con el ojo?

Table 4: Examples of parallel questions from the English/German and the English/Spanish datasets.

[7] https://sites.google.com/site/marianalaraneves/resources/

3. System architecture

Question answering systems are usually composed of three steps (Athenikos and Han, 2010): question processing, passage retrieval and answer processing. In the first step, the system identifies the type of question which has been posed (e.g., yes/no, definitional or factoid), the expected answer e.g., gene/protein, disease, etc.), in case of factoid questions, and converts the question into a query to the passage retrieval step. It might also include identification of semantic concepts and query expansion based on available lexical thesaurus (e.g. WordNet) or domain-specific resources (e.g., UMLS). In the passage retrieval step, queries are posed to a collection of documents and passages, usually a couple of sentences, are ranked according to their relevancy to the query.

In this section, we describe the QA system prototype that is under development and that has been used for a preliminary evaluation of the parallel collection of questions. For this work, we focus on the question processing and passage retrieval steps, as we do still do not provide suggestions of answers for the proposed questions.

3.1. Question processing

Our prototype system starts with the tokenization of the question followed by the part-of-speech tagging of resulting tokens. We use the OpenNLP Maximum Entropy-based tokenizer and part-of-speech tagger and the corresponding available models for English and German[8]. Given that no models are available for the Spanish language, we have used the one available for Portuguese for the tokenization step, given the similarity between these languages and that tokenization is even more challenging in the later due to the composed words, which are not common in Spanish. For the part-of-speech tagging in Spanish, we have used the Maxent model made available by Juan Manuel Caicedo Carvajal[9].

Some tokens were filtered out according to language-specific Stopwords lists and to the part-of-speech tags which indicates numerals, e.g., "CD" for English, "CARD" for German, "DN" and "Z" for Spanish. For the later, we have used existing lists available for English[10], German and Spanish (both from the stop-words project[11]) and we have extended them occasionally with extra words which were missing, such as the prepositions "de" and "al" from the Spanish.

We carry out a query expansion of the tokens using the thesaurus made available in the CLEF-ER challenge (cf. Section 2.1.). The CLEF-ER thesaurus has been loaded into the HANA database and a fuzzy search is carried out for each token in the query against the synonyms available for the corresponding language. Comparison of the query term and the synonyms is performed by requiring at least 90% similarity of the terms, to allow more flexibility of the matching. For performance reasons and in order not to include potential irrelevant synonyms, we only expanded the query with the synonyms identified as preferred (e.g, "PREF_EN", "PREF_DE") in the CLEF-ER terminology (cf. Table 2).

Weights are assigned for the terms of the query and the corresponding synonyms and are calculated based on the popularity of the term in the CLEF-ER terminology. The higher the number of synonyms which match to a term, the lower the weight of the later. Tokens which do not match any synonyms are assigned weight 0.5, i.e., an average weight. Otherwise, weights are calculated based on the number of terms which matched to this particular token (#MatchesToken) and the total number of terms matched to all tokens of the query (#MatchesTotal), following the expression below:

$$weight = 1 - \frac{\#MatchesToken}{\#MatchesTotal} \quad (1)$$

3.2. Passage retrieval

Passage retrieval is performed using the SAP HANA database[12] (hereafter called HANA), an in-memory database which has already been successfully employed for real-time analysis of biomedical data (Schapranow et al., 2013). HANA provides built-in text analysis functionalities for eleven languages (Arabic, simplified Chinese, Dutch, English, Farsi, French, German, Italian, Japanese, Korean, Portuguese, Spanish) which includes integrated sentence splitting, tokenization and stemming components during full text indexing of the documents. Further, it allows fuzzy searching the full text index according to a pre-defined percentage (e.g., 90%) and also linguistic searching by considering linguistic variants of the query terms.

The Medline documents available for the English, Spanish and German languages from the CLEF-ER (cf. Section 2.1.) were loaded into HANA and a full text index was created for each collection. We have experimented with three search strategies provided by HANA: exact, fuzzy matching (at least 90% of similarity) and linguistic. When considering the linguistic search, which allows matching of terms which share the same stem, we apply it only to the terms of the query which have not been matched to any synonym. When querying for relevant passages, HANA proceeds in matching the terms to the individual tokens derived from the full text index of the documents according to the specified search strategy, whether exact, fuzzy or linguistic. A score is calculated by HANA for the matching tokens and sentences are ranked according to the weighted sum of scores of all matching tokens per sentence. As HANA automatically split the sentence during indexing of

[8] http://opennlp.sourceforge.net/models-1.5/

[9] http://cavorite.com/labs/nlp/opennlp-models-es/

[10] http://www.textfixer.com/resources/common-english-words.txt

[11] https://code.google.com/p/stop-words/

[12] http://www.saphana.com/

the documents, the current system could also be applied to abstracts of full text documents and not only to titles (usually single sentences). Table 5 shows the top ten sentences retrieved for the one of the English questions shown in Table 4.

4. Experiments and results

We randomly split the two parallel collections of 50 questions in English/Spanish and English/German in two sets of 25 question each, i.e., two sets of 25 questions for development and two sets of 25 questions for testing purposes. The development datasets were used for choosing the best strategies, adding of extra stopwords and error analysis, but have not been used to train any of components of the system.

Evaluation of the development and test datasets consisted in running the system for each of them, retrieving the 10 best ranked passages (sentences) and checking whether the original document from which the question had been derived was present in this list. With such an experiment, we sought to evaluate the precision of our QA prototype for finding relevant passages to the questions as well as the difficulty level of the questions. Text passages were retrieved only for documents in the same language of the question, e.g., Spanish questions were queried only against the Spanish documents, thus, no machine translation was used.

We have evaluated the following settings of our QA system:

1. HANA exact search;
2. HANA fuzzy matching (at least 90% similarity);
3. HANA fuzzy matching (above), plus query expansion of the question words using the CLEF-ER terminology;
4. HANA fuzzy matching (above), query expansion (above), plus HANA linguistic matching for those words in the question which did not match to any synonym.

For the evaluation, we calculate the R-Precision, which is the precision on the r-th position where a first match with the original document was found, or zero, if not found. For instance, if the first match is found in the third position, the R-P is 0.33 (1/3). We then calculate the mean of the R-precision over the collection of questions, i.e., over 25 questions for each dataset. Table 6 shows the results for each language in the English/German and English/Spanish parallel collections of questions.

Settings	English-German		English-Spanish	
	EN	DE	EN	ES
exact	0.04 (1)	0.04 (1)	0.01 (1)	0.09 (4)
fuzzy	0.05 (3)	0.02 (1)	0.10 (5)	0.10 (5)
+ synonyms	0.09 (3)	0.06 (2)	0.11 (4)	0.09 (6)
+ linguistic	0.10 (4)	0.04 (2)	0.03 (2)	0.08 (6)
Test	0.05 (5)	0.05 (4)	0.05 (3)	0.06 (4)

Table 6: R-Precision and number of sentences found (in parenthesis) for each dataset and for various setting of the question answering system. Results for the training dataset are shown for the many setting options while the ones for the test dataset were obtained when using query expansion and fuzzy matching (i.e., "+ synonyms"). Codes for the languages are the following: "EN": English, "DE": German, "ES": Spanish.

5. Discussion and future works

In this work, we have presented a preliminary evaluation of a passage retrieval component for a multilingual question answering system on two datasets of 50 parallel questions for English/German and English/Spanish. Regarding the construction of the question collection, we believe that 50 questions per collection (100 in total) is an adequate number for evaluation purposes, given that previous challenges in this field have utilized datasets with similar size, such as the 40 questions from the in the machine reading dataset for Alzheimer Disease (Morante et al., 2012) and the batches of 100 questions released during the BioAsq challenge (Partalas et al., 2013). As described in Section 2.2., the selection of the questions was carried using a hybrid approach of first randomly retrieving batches of 50 document titles from the CLEF-ER Medline collection and then manually choosing those which were more related to the biological domain and in order to avoid irrelevant titles. We define a relevant title as those that contain enough information to build an interesting factoid question without the need to refer to the abstract or the full paper of the publication. We believe that this approach ensures both the variability of our dataset, through the random selection of the candidates, as well as its quality through a subsequent manual selection of relevant titles.

Deriving questions from the original document titles required deciding which entity type, whether a species or a disease, was particularly interesting to be the subject of the answer. Writing the question collection was a task that took from 5 to 10 minutes per question, depending on the difficulty in rephrasing the text and in finding appropriate synonyms for the biological terms and remaining words, which required authors to refer to several on-line resources. For instance, for a certain question, "polio" has been used as synonym to "poliomyelitis" and "factor favoring" has been rephrased as "cause increased severity". Acronyms were frequently used whenever available, such as "HCV" instead of "hepatitis C virus". In some cases where no adequate synonym could be found for a term, we have opted for using hypernyms instead, such as "fish" in substitution for "tilapia", which requires question answering systems to consider hierarchical relationships between terms. Other types of relationships have been also explored, such as referring to a organ instead of a disease name, for instance, "eye related infections" instead of "keratoconjunctivitis".

Which methods can be used to determine the living cell count of cariogenic microorganisms?
1. Determination of the living cell count of cariogenic microorganisms using the measurement of their ATP content in the bioluminescence procedure–a critical look at the method. 2. Comparison of the coulter-counter-method and the counting-chamber to determine the cell count of the cerebrospinal fluid. 3. Phase contrast microscopic studies of living germs, a supplemental method for the study of effect of germicidal substances on microorganisms. 4. Germ count determination" from water samples by the plate method with special reference to the pH value of the used nutrient media. 5. Schistocytes: which definition should be taken and which method should be used to identify and count them?. 6. Comparative survey concerning methods of vestibular exploration used in 10 Western European countries. 7. The methods used to collect hematopoietic stem cells. 8. Comparative evaluation of the methods used to determine the sensitivity of bacteria to antibiotics. 9. Critical study of various methods used to determine the activity of anti-typhoid vaccine. 10. Human bone marrow cell culture–a sensitive method for the evaluation of the biocompatibility of materials used in orthopedics.

Table 5: Retrieved passages for a English question. The original title from which the question has been derived is the one in the first position of the rank.

In general, preliminary results show that passage retrieval is not a straightforward task for none of the three languages, even when using a small collection of Medline document titles. Although the best results were obtained for Spanish, this was probably due to the smaller set of documents available for this language (cf. Table 3) than to the simplicity of the task or the language. On the other hand, results for German were particularly low in comparison to the other languages because of the presence of many compound words, as will be discussed below. Despite the overall low performance of our results, these experiments provide baseline results for the proposed collection and gives an estimation on the quality of our dataset.

In comparison to the development dataset, performance for the test dataset was higher for the English/German dataset and a little bit lower for the English/Spanish questions. Curiously, this is the best performing result for the German questions. Discrepancies between development and test data are expected given the small number of questions and the distinct topics they refer. The development dataset was used only for comparing the different setting of the features, updating the stopwords list with additional terms and performing an analysis of the results.

Results vary significantly for the many configurations of our system when evaluated for the three languages (cf. Table 6). Despite the less complexity of the English language, more exact matches have been obtained for the Spanish (4) than for English (1 for each dataset) or German (1). The higher number of matches for Spanish have occurred mainly due the impossibility in getting better synonyms in the Spanish language when writing the question, together with the fact that less documents are available for this language. For instance, the question "Qué métodos histoquímicos permiten la observación y caracterización de glicoconjugados?" got the right match (document d9279022) in the third rank, while no correct match was found for the corresponding English question ("Which histochemical methods allow observation and characterisation of glycoconjugates?").

As expected, consideration of fuzzy matching (at least 90% of similarity) improves both the average R-Precision and the number of retrieved documents for all languages (except for German), as more terms can be matched using this approximate matching. Additionally, this improvement came with no degradation of the R-Precision, which also increased for all languages, again, with the exception of German. For the later, the only question which got a match in the first rank when using exact matching, "Spielt die Methode der RNA in situ Hybridisierung bei Studien an Rhesusaffen eine Rolle?" (document 7534003), this time got a match in the second rank, thus the decrease in precision. This was due to the fuzzy match between the tokens "HIV-1-RNA" and "RNA", as words are always tokenized by the dashes in HANA.

Contrary to the expected, the query expansion brought just one additional match for the German and Spanish datasets, but also one less for one of the English datasets, while not changing the other one. On the other hand, it increased the average R-Precision for most of the datasets, except for the Spanish one. This improvement occurred also because of the weights associated to each term, which help giving higher score to sentences containing terms with associated synonyms, as opposed to common words of the language not related to biomedical domain. As an example of an improvement due to query expansion, the short Spanish question "Qué patógenos de mosquitos existen?" ("Which mosquito pathogens exist?") got no matches in the fuzzy search but got one when using query expansion due to the addition of the synonym "Culicidae". The same synonym was retrieved for the corresponding English question but the large collection of documents make it more difficult to get the correct document in the top results for such a general question. However, many the documents which were returned to this English questions could also be considered as correct, only that none of them was the one originally used for creating the question. Thus, future work could include the expansion of the questions datasets with other relevant passages (document titles) other than the original one.

Nevertheless, query expansion did not help much in getting new terms due to the limitations of the CLEF-ER terminology which is based only on three resources (cf. Section 2.1.) and without the exploration of the relationships between the terms, i.e., hypernyms and hyponyms. For instance, the adjective "corneal" was not obtained for the term "eye", although the synonym "Corneal Disease" does exist in the terminology. Another example of synonyms which could not be found only relying on the CLEF-ER thesaurus is the pair "endovenous" and "intravenous", none of them is present in the resource. Further, the use of a token-based query expansion did not allow the correcting matching of the terms, along with the complexity of the biomedical nomenclature. For instance, it is not straightforward to match "Staphylococcus aureus" to "S. aureus" without consideration of pattern rules specific for the species nomenclature.

Finally, we studied the use of HANA linguistic search for those terms which did not match synonyms in the CLEF-ER terminology, which we expected to be common words of the language instead and named-entities. However, this search approach did not get additional document matches due to the way that the score is calculated by HANA when using this type of search. The only additional match we got for an English document seems not to be directly related to this feature.

An analysis of the questions to which the corresponding documents could not be found in the top 10 ranking results gives us insights on the next steps to improve our system. First of all, different strategies might be necessary for different languages. While token-based search successfully obtained some matches for English and Spanish, it failed to perform well for German due to the high number of compound words. For instance, if a question contains the word "Pankreaskarzinom" and the document the term "Adenokarzinom" neither an exact nor 90% fuzzy matching would be able to get such match, while the word "carcinom" (in English) and "carcinoma" (in Spanish) would match in the corresponding questions and documents in these languages. For future work, we want to explore advanced natural language processing functionalities of the HANA database which identifies compound words for German. For instance, the word "hochdosiert" (high dosed), which is an adjective derived from the adjective "hoch" (high) and the verb "dosieren" (to dose), is identified in the HANA full text index as "hoch#dosieren", thus allowing partial matches to any of the words in its composition.

We have relied in external libraries (i.e., OpenNLP) for the pre-processing of the questions, which has been also limited to the models available for the corresponding languages. As future work, we will study the use of the HANA database for this step as well, given that it already provides support for these languages as well as shallow parsing information (chunks). The later has not been explored in this work but can certainly help in both the query expansion, when matching query terms to potential synonyms, as well as in the passage retrieval step, when matching terms to documents. Further, other metrics for the terms weights and a range of values for the fuzzy matching will also be studied.

Regarding the semantic resources, future works will certainly use additional multilingual terminologies, such as DBpedia[13], as well as the exploration of semantic relationships between the terms for obtaining also hypernyms and hyponyms. Also additional lexical resources will be explored for retrieving synonyms and related words to common words of the languages, such as Wordnet for English, and similar resources for the other languages. In this work, the background collection has been limited to the document titles made available during the CLEF-ER challenge, given the scarce resources of biomedical documents in languages other than English. Therefore, future work could also explore the use of machine translation (Jimeno Yepes et al., 2013) for translating question in other languages to English, thus being able to query the English documents in Medline.

In this work, we have limited languages to English, German and Spanish according to the languages which the authors were more confident, as well as the ones supported by both the CLEF-ER resources and the HANA database. However, changes in the systems for any of the other languages already supported by HANA (Arabic, simplified Chinese, Dutch, Farsi, French, Italian, Japanese, Korean, Portuguese) would only require a background collection of documents and a pertinent list of questions.

6. Conclusions

In this work we have presented our prototype of a multilingual question answering system for the biomedical domain and have evaluated the system using two collections of 50 questions for English/German and English/Spanish. Our system works on the top of a SAP HANA database which includes built-in text analysis functionalities, such as quick indexing of large collections of documents, embedded natural language processing (sentence splitting, tokenization and shallow parsing) and querying the text collection using weighted fuzzy and linguistic matching. Our evaluation for the passage retrieval task has shown the particularities of each language and has pointed out which resources are necessary for each of them in order to boost multilingual question answering results for the biomedical domain.

7. Acknowledgements

We would like to thank our colleagues in the Hasso-Plattner Institut (HPI), specially Dr. Matthieu-P. Schapra-

[13] `http://dbpedia.org/`

now, Cindy Fähnrich and Thomas Uhde. MN would like to acknowledge funding from the HPI Research School.

8. References

Sofia J. Athenikos and Hyoil Han. 2010. Biomedical question answering: A survey. *Computer Methods and Programs in Biomedicine*, 99(1):1 – 24.

MichaelA Bauer and Daniel Berleant. 2012. Usability survey of biomedical question answering systems. *Human Genomics*, 6(1):1–4.

Michael Heilman and Noah A. Smith. 2010. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 609–617, Stroudsburg, PA, USA. Association for Computational Linguistics.

William Hersh and Ellen Voorhees. 2009. Trec genomics special issue overview. *Inf. Retr.*, 12(1):1–15, February.

Antonio Jimeno Yepes, Elise Prieur-Gaston, and Aurelie Neveol. 2013. Combining medline and publisher data to create parallel corpora for the automatic translation of biomedical text. *BMC Bioinformatics*, 14(1):146.

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 28–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Roser Morante, Martin Krallinger, Alfonso Valencia, and Walter Daelemans. 2012. Machine reading of biomedical texts about alzheimer's disease. In *CLEF (Online Working Notes/Labs/Workshop)*.

Mara-Dolores Olvera-Lobo and Juncal Gutirrez-Artacho. 2011. Multilingual question-answering system in biomedical domain on the web: An evaluation. In Pamela Forner, Julio Gonzalo, Jaana Keklinen, Mounia Lalmas, and Maarten de Rijke, editors, *CLEF*, volume 6941 of *Lecture Notes in Computer Science*, pages 83–88. Springer.

Ioannis Partalas, Eric Gaussier, and Axel-Cyrille Ngonga Ngomo. 2013. Results of the first bioasq workshop. In *1st Workshop on Bio-Medical Semantic Indexing and Question Answering, a Post-Conference Workshop of Conference and Labs of the Evaluation Forum 2013 (CLEF 2013)*, nov.

Dietrich Rebholz-Schuhmann, Simon Clematide, Fabio Rinaldi, Erik Van Mulligen, Ian Lewin, David Milward, Antonio Jimeno Yepes, Udo Hahn, Jan Kors, Chinh Bui, Johannes Hellrich, and Michael Poprat. 2013. Multilingual semantic resources and parallel corpora in the biomedical domain: the clef-er challenge. In *CLEF Lab*, September.

Matthieu-P. Schapranow, Hasso Plattner, and Christoph Meinel. 2013. Applied in-memory technology for high-throughput genome data processing and real-time analysis. In *System on Chip (SoC) Devices in Telemedicine from LABoC to High Resolution Images, pp. 35-42, ISBN: 978-84-615-6080-6.*