Machine Reading of Biomedical Texts about Alzheimer's Disease

[1]Roser Morante, [2]Martin Krallinger, [2]Alfonso Valencia, and [1]Walter Daelemans

[1]CLiPS, University of Antwerp, Prinsstraat 13, B-2000 Antwerpen, Belgium
{roser.morante,walter.daelemans}@ua.ac.be
[2]CNIO, Melchor Fernández Almagro 3, 28029 Madrid, Spain
{mkrallinger,avalencia}@cnio.es

Abstract. This report describes the task *Machine reading of biomedical texts about Alzheimer's disease*, which is a pilot task of the Question Answering for Machine Reading Evaluation (QA4MRE) Lab at CLEF 2012. The task aims at exploring the ability of a machine reading system to answer questions about a scientific topic, namely Alzheimer's disease. As in the QA4MRE task, participant systems were asked to read a document and identify the answers to a set of questions about information that is stated or implied in the text. A background collection was provided for systems to acquire background knowledge. The background collection is a corpus newly compiled for this task, the Alzheimer's Disease Literature Corpus. Seven teams participated in the task submitting a total of 43 runs. The highest score obtained by a team was 0.55 c@1, which is clearly above baseline.

1 Introduction

This report describes the task *Machine reading of biomedical texts about Alzheimer 's disease*, which is a pilot task of the *Question Answering for Machine Reading Evaluation* (QA4MRE)[1] Lab at CLEF 2012. The task follows the same set up and principles as the QA4MRE task (7; 8), with the difference that it focuses on the biomedical domain.

This pilot task aims at exploring the ability of a machine reading system (4; 12) to answer questions about a scientific topic, namely the Alzheimer's disease (AD). AD has been chosen as the focus of the task because there is a particular interest in more efficient processing of Alzheimer-related literature, as this condition constitutes a considerable health challenge for an aging population (Citron 2010). The increasing importance of AD is reflected in the recently approved US National Alzheimer's Project Act,[2] which will result in considerable funding being made available for research on this disease and for financing better data infrastructure resources. Currently, the illness is being analyzed from various perspectives in a growing number of scientific studies (5; 1; 2).

[1] http://celct.fbk.eu/QA4MRE/
[2] http://aspe.hhs.gov/daltcp/napa/#NAPA

The report is organised as follows. The task is described in Section 2, Section 3 provides information about the Alzheimer's Disease Literature Corpus and Section 4 about the test data. Section 5 explains the process followed to annotated the data. Section 6 deals with the design of questions. In Section 7 the evaluation process is explained and in Section 8 details about the number of participating systems and runs are presented as well as their results. Finally, Section 9 closes the paper with some conclusions.

2 Task description

As in the QA4MRE task, participant systems were asked to read a document and identify the answers to a set of questions about information that is stated or implied in the text. Questions are in the form of multiple choice, each having five options, and only one correct answer. The detection of correct answers is specifically designed to require various kinds of inference and the consideration of previously acquired background knowledge. Knowledge acquisition can be performed from a document collection called the *background collection* provided by the organization. Although the additional knowledge obtained through the background collection may be used to assist with answering the questions, the principal answer is to be found among the facts contained in the test documents. Participants were provided with a collection of texts about Alzheimer's Disease called the Alzheimer's Disease Literature Corpus (ADLC corpus), which was compiled for this task. The evaluation is performed on four reading tests with ten multiple choice questions each.

To solve the task, participants could make use of existing resources, such as ontologies or databases, and tools, such as named entity taggers, event extractors, parsers, etc. In order to keep the task reasonably simple for systems, the task organizers provided the texts of the background collection and the test documents processed at several levels of linguistic analysis (lemmas, part-of-speech, named entities, chunking, dependency parsing) with publicly available state of the art tools.

3 Background collection: the Alzheimer's Disease Literature Corpus

The background collection is a collection of texts about Alzheimer's disease called the Alzheimer's Disease Literature Corpus (ADLC corpus). Systems could use it to acquire reading capabilities and to obtain knowledge about Alzheimer's disease that could help in answering the questions about the test documents. The texts have been carefully selected to be as specific as possible for this topic and the corpus should constitute a comprehensive resource for this task in particular and for text mining efforts tailored to the Alzheimer's disease field in general. Although the use of the background collection is recommended, it is not

mandatory. The background collection is released subject to signing a license agreement.[3] It contains the following sets of documents:

PubMed abstracts. 66,222 abstracts obtained by performing in PubMed the search provided in Figure 1. The abstracts were provided in XML format, and with the annotations described in Section 5.

((((((”Alzheimer Disease”[Mesh] OR ”Alzheimer’s disease antigen”[Supplementary Concept] OR ”APP protein, human”[Supplementary Concept] OR ”PSEN2 protein, human”[Supplementary Concept] OR ”PSEN1 protein, human”[Supplementary Concept]) OR ”Amyloid beta-Peptides”[Mesh]) OR ”donepezil”[Supplementary Concept]) OR (”gamma-secretase activating protein, human”[Supplementary Concept] OR ”gamma-secretase activating protein, mouse”[Supplementary Concept])) OR ”amyloid beta-protein (1-42)”[Supplementary Concept]) OR ”Presenilins”[Mesh]) OR ”Neurofibrillary Tangles”[Mesh] OR ”Alzheimer’s disease”[All Fields] OR ”Alzheimer’s Disease”[All Fields] OR ”Alzheimer s disease”[All Fields] OR ”Alzheimers disease”[All Fields] OR ”Alzheimer’s dementia”[All Fields] OR ”Alzheimer dementia”[All Fields] OR ”Alzheimer-type dementia”[All Fields] NOT ”non-Alzheimer”[All Fields] NOT (”non-AD”[All Fields] AND ”dementia”[All Fields]) AND (hasabstract[text] AND English[lang])

Fig. 1. Search performed in PubMed.

Open Access full articles PMC. 8,249 Open Access full articles from PubMed Central in PDF format. These articles have been selected by first performing the search in Figure 1 and then selecting the full articles that belong to the PubMed Central Open Access subset and that were available on 1.03.2012. 7,512 of these articles were provided in text format, which was obtained by converting the PDF files into text by using the tool LA-PDFText[4] (9). 7,447 of these articles were also provided with annotations.

Open Access full articles PMC, smaller set. This smaller set contains 1,041 full text articles from PubMed Central in HTML and text format. The articles are also provided with annotations. For this articles the text version has been converted from the PubMed HTML version. To select these documents a search was performed on PubMed using Alzheimer’s disease related keywords and restricting the search to the last three years. The search was performed on 3.02.2012. Only a subset of the articles obtained by the search has been included in the collection.

Elsevier full articles. This set contains 379 full text articles from Elsevier and 103 abstracts. The documents are provided in XML and text format. They are also provided with annotations. The text files have been obtained by converting the XML files into text. The articles in this subset have been

[3] The ADLC corpus can be downloaded from the following link: `http://celct.fbk.eu/ResPubliQA/index.php?page=Pages/bg_collection_pilot.php`

[4] LA-PDFText is available at `http://code.google.com/p/lapdftext/`

selected from a list of articles provided by Professor Tim Clark from the Massachusetts Alzheimer's Disease Research Center, USA. The list contains bibliographic records representing 45 core hypotheses in Alzheimer's disease. Elsevier kindly provided the articles from this list that were Elsevier publications.

4 Test data

The test set is composed of 4 reading tests, each consisting of 10 questions about 1 document, with 5 answer choices per question. So, there were in total 40 questions and 200 choices/options. Participating systems were required to answer these 40 questions by choosing in each case one answer from the five alternatives. Systems could leave questions unanswered.

The test documents were selected from a list of bibliographic records provided by professor Tim Clark from the Massachusetts Alzheimer's Disease Research Center, USA. The records were compiled in 2011 and represent 45 core hypotheses in Alzheimer's disease.

The test documents were provided in text format. They were first converted automatically from PDF into text format and then the text version was corrected manually, paying attention to symbols that express relevant information about Alzheimer's disease. The captions of figures and tables were also included, but the figures and tables not. Participants were not expected to process the contents of tables and figures. A sample of a test document with questions can be downloaded from the QA4MRE website.[5] The test documents and the questions were provided also with annotations.

5 Data annotation

The documents in the background collection, the test documents, and the questions were provided with annotations in a column format as shown in Figure 2.

22319430	1	1	HD	HD	B-NP	NN	B-protein	4	NMOD	B-XHD_amino_acid_duplex	B-PROTEIN
22319430	1	2	amino	amino	I-NP	JJ	I-protein	4	NMOD	I-XHD_amino_acid_duplex	I-PROTEIN
22319430	1	3	acid	acid	I-NP	NN	I-protein	4	NMOD	I-XHD_amino_acid_duplex	I-PROTEIN
22319430	1	4	duplex	duplex	I-NP	NN	I-protein	5	SUB	I-XHD_amino_acid_duplex	I-PROTEIN
22319430	1	5	has	have	B-VP	VBZ	O	0	ROOT	O	O
22319430	1	6	been	be	I-VP	VBN	O	5	VC	O	O
22319430	1	7	found	find	I-VP	VBN	O	6	VC	O	O
22319430	1	8	in	in	B-PP	IN	O	7	VMOD	O	O
22319430	1	9	the	the	B-NP	DT	O	11	NMOD	O	O
22319430	1	10	active	active	I-NP	JJ	O	11	NMOD	O	O
22319430	1	11	center	center	I-NP	NN	O	8	PMOD	O	O
22319430	1	12	of	of	B-PP	IN	O	11	NMOD	O	O
22319430	1	13	many	many	B-NP	JJ	O	12	PMOD	B-Xmany_different_enzyme	O
22319430	1	14	different	different	I-NP	JJ	O	15	NMOD	I-Xmany_different_enzyme	O
22319430	1	15	enzymes	enzyme	I-NP	NNS	B-protein	13	NMOD	I-Xmany_different_enzyme	B-PROTEIN
22319430	1	16	.	.	O	.	O	5	P	O	O

Fig. 2. Example of an annotated sentence.

[5] http://celct.fbk.eu/QA4MRE/index.php?page=Pages/downloads.php

The annotations were obtained automatically with the dependency parser GDep (10), a UMLS (3) based NE tagger developed at CLiPS, and the ABNER NE tagger (11). The content of the columns is specified in Table 1.

Column 1	Document identifier
Column 2	Sentence number in the document
Column 3	Token number in the sentence
Column 4	Word (GDep parser)
Column 5	Lemma (GDep parser)
Column 6	Chunk tag (GDep parser)
Column 7	Part-of-speech tag (GDep parser)
Column 8	Named entity (GDep parser)
Column 9	Parent node in the dependency syntaxt tree (GDep parser)
Column 10	Dependency syntax label (GDep parser)
Column 11	UMLS named entity (CLiPS NE Tagger)
Column 12	Named entity (ABNER tagger)

Table 1. Annotated information.

6 Question design

As in the QA4MRE task, questions are in multiple choice format and focus on testing the comprehension of one single document. The questions posed for this task should address aspects that are of biomedical relevance and that have been proven to be of importance in the context of previous efforts such as BioCreative[6], Genomics TREC track[7] or the BioNLP shared tasks.[8] This should enable participants to make use of resources developed for these competitions and will establish a link between this pilot task and previous efforts. Additionally, since machine reading of biomedical texts is a new task, it seemed more appropriate to restrict the types of questions somehow. Therefore a restricted set of named entity types associated to the questions was defined, as well as a list of question types. The expected answer types for the multiple choice answers depend on allowed entity types.

6.1 Named entities

The categories of named entities considered for this task are the following:

- GENE_PROT. Genes and gene products (proteins, mRNA).
- CHEM_DRUG. Chemicals/drugs/pharmacological agents.
- DIS_SYMPT. Disease/symptoms.

[6] `http://www.biocreative.org`
[7] `http://ir.ohsu.edu/genomics`
[8] `http://sites.google.com/site/bionlpst`

- EXP_METHOD. Experimental method/qualifier.
- SPEC_ORG. Species/organism.
- PATH_PROC. Pathway/Biological process.
- ANAT_CELL. Anatomical/cellular/subcellular structures.
- MUT_PTM. Mutations/genetic variations/posttranslational modifications.
- ADV_TOXIC. Adverse effect/toxic endpoints.
- DOSE. Dose of a given treatment.
- TIMING. Schedule of treatments (timing).
- PAT_CHAR. Patient characteristics: age, gender, sex, race, population, animal strain.
- MOL_MARKER. Molecular marker.

In order to identify the named entities above, the following lexico-semantic resources and tools can be used (among others): ABNER, BANNER, Genia Tagger, BioThesaurus, BioLexicon,UMLS, LINNAEUS tagger, OrganismTagger, MeSH, Gene Ontology (and other ontologies from OBO), etc... .

The test documents were processed with UMLS and the BANNER tagger before making the questions, so that questions would refer only to entities that can be automatically identified with existing resources.

6.2 Question types

Based on examination of the relationships between the various entity types we compiled the following collection of biomedically relevant question types:

Experimental evidence/qualifier. This question type refers to experimental techniques, methods or models used to generate or validate a given discovery. Examples include animal models used for a given in vivo study, interaction detection methods used to detect protein interactions, imaging techniques for visualization or localization of a particular protein.

Protein-protein interaction. This question type refers to the detection of an interaction partner of a given protein. Examples include physical binding of two proteins in a protein-protein complex or more transient interaction in phosphorylation of one protein by another.

Gene synonymy relation. This question type tries to establish relations between two entity mentions of genes or proteins that refer actually to the same biological entity. For instance this relation exists between 'APP' and 'amyloid beta (A4) precursor protein'. Here alternative aliases of a gene name or symbol are included, as well as typographical variants and acronyms and their corresponding expanded forms.

Organism source relation. This question type refers to the actual organism source for a given protein or gene. An example would be the genes encoded in the human genome or expressed in humans.

Regulatory relation. This question type refers to gene regulatory relationships between two bio-entities (protein and gene), i.e. whether one bio-entity affects the gene expression of another entity (e.g. transcription factor target gene relation).

Increase (improvement, higher expression). This is a more specific question type of the regulatory relation. It refers to cases where one bio-entity causes the upregulation (increased expression) of another bio-entity.

Decrease (depletion, reduction). This is a more specific question type of the regulatory relation. It refers to cases where one bio-entity causes the downregulation (decreased expression) of another bio-entity.

Inhibition/disruption/impaired. This question type refers to cases were one bio-entity blocks or inhibits another bio-entity. Examples include drugs blocking a given protein or enzyme, or proteins that inhibit a particular biological process or pathway.

We found that some types of questions defined initially based on an example test document did not occur frequently in the document tests used for evaluation. For a future edition of the task we would like to select entity types that are guaranteed to occur in all documents. Document selection could also be performed depending on the most frequent entity types.

6.3 Degrees of difficulty

Questions can be assigned a degree of difficulty: simple, medium and complex.

Simple. Factual questions that can be answered using information from the target document and whose textual evidence is contained multiple times in the paper, e.g. several text snippets are supporting the correct answer. The answer is found almost verbatim in the paper.

Medium. The correct answer is phrased in a way that requires the use of lexico-semantic dictionaries and name alias recognition capabilities to be able to handle lexico-semantic alienations of keywords and entities.

Complex. Reasoning must be applied to answer this question. Choosing the correct answer requires combining pieces of evidence. Such questions might need ad hoc axiomatic knowledge and abductive processes.

A collection of criteria for question difficulty classification was followed. Aspects that influence question difficulty include:

- Are the ontological relations encoded in the question? If they are encoded the question should be easier.
- If keyword-based indexing and conceptual indexing are required the question is less easy.
- Script like questions such as 'how is an anatomical structure assembled?' should be more difficult since answering them requires combining several units of information.
- Template questions about successive temporal events (biological processes, disease stages) should be more difficult since it also requires several units of information.

- Is it necessary to process morphological alternations such as *phosphorylate* lexicalized as the nominalization *phosphorylation*? In this case the degree of difficulty should be simple/medium, depending on other characteristics of the question.
- Is it necessary to process lexical alternations? The usage of synonyms or semantically related terms derived from ontologies is necessary to increase the recall.
- Is it necessary to process semantic alternations and paraphrases? This involves finding relations between multi-term paraphrases and single terms, textual patterns, or complex examination between word building terms within the ontology.
- Is it necessary to process terminological variants and high level indexes comprising terms and their variants for retrieval? A variant recognition module is required as well as weighting of matching between questions and documents.
- How big is the paragraph window size of the evidence text? Is it a continuous span of text? The bigger the window size, the more difficult is the question. Non continuous spans are more difficult to process than continuous.

Assigning degrees of difficulty to the questions proved to be a difficult task because several factors interact. For a future edition of the task we would like to define a protocol in order to facilitate the assignment of degrees of difficulty, which can also be used to perform error analysis.

6.4 Answers

As in the main task, systems are not required to answer every question, since the c@1 measure (6) was used for evaluation. This measure encourages systems to reduce the number of incorrect answers while maintaining the number of correct ones by leaving some questions unanswered. Systems were asked to choose the right answer among five choices.

7 Evaluation

As in the main task, participants were allowed to submit a maximum of 10 runs. Each run should be categorized as one of the following types, depending on the resources that have been used to assist in asnwering the questions:

1. No external resource was used (only the test document).
2. Only the test document and the associated background collection was used.
3. The test document and other resources were used, but not the background collection.
4. The test document together with the background collection and other resources were used.

Evaluation was performed automatically following the same procedure as in the QA4MRE task. Each question received one (and only one) of the three following assessments:

- *Correct* if the system selected the correct answer among the five candidate ones of the given question.
- *Incorrect* if the system selected one of the wrong answers.
- *NoA* if the system chose not to answer the question.

The main evaluation measure used was c@1 (6), which takes into account the option of not answering certain questions. The formulation of c@1 is given in (1). The overall c@1 is calculated over the 40 questions of the test collection.

$$c@1 = \frac{1}{n}(n_R + n_U \frac{n_R}{n}) \tag{1}$$

where

n_R: number of questions correctly answered.
n_U: number of questions unanswered
n: total number of questions

As a secondary measure systems are evaluated on accuracy, which is the traditional measure applied to question answering evaluations that do not distinguish between answered and unanswered questions. The formulation of accuracy is given in (2). The overall accuracy is calculated over the 40 questions of the test collection.

$$accuracy = \frac{n_R + n_{UR}}{n} \tag{2}$$

where

n_R: number of questions correctly answered.
n_{UR}: number of unanswered questions whose candidate answer was correct.
n: total number of questions

More information about the evaluation procedure can be found in the article that describes the QA4MRE task (8).

8 Participation and results

Out of the 23 groups that had previously registered and signed the license agreement to download the background collection, a total of 7 groups participated submitting 43 runs. Table 2 shows the list of participating teams and the reference to their reports.

Table 3 provides information about the number of runs per team and the scores of the best run in terms of c@1. A random baseline is calculated, assuming that a system answers all questions. This baseline has five possibilities when trying to answer a question: it can select the correct answer to the question, or it can select one of the four incorrect answers. In this case, the overall result is 0,20. One of the participating systems scores as the baseline, whereas the team

Team	Institutions	Reference
lims	LIMSI-CNRS - Université Paris-Sud - ENSIIE, France – Fondazione Bruno Kessler, Italy	Grau et al.
kule	Katholieke Universiteit Leuven, Belgium	Verbeke and Davis
ntnu	National Taiwan Normal University, Taiwan	Tsai et al.
merk	The University of Iowa, USA – Merck KGaA, Germany	Bhattacharya and Toldo
iirg	University College Dublin, Ireland	Byrne et al.
Pisa	Università di Pisa, Italy	Attardi et al.
nict	National ICT Australia - Macquarie University, Australia	Martínez et al.

Table 2. Participating teams with reference to their reports.

that obtained the best results is clearly above baseline. Given the fact that four of the seven teams are only some points above baseline, we consider that the task is complex and the questions put were not easy to answer. However, given that two systems are clearly above baseline, we consider that the difficulty of the questions provided was suitable for the task.

Team	# of runs	Highest c@1 score
Pisa	1	0,55
merk	7	0,47
kule	10	0,30
nict	9	0,28
iirg	7	0,25
lims	4	0,21
ntnu	5	0,20
baseline	-	0,20

Table 3. Number of runs and highest scores per team.

The team that obtained the highest scores, *Pisa*, approaches the task using the index expansion technique, adding variants of terms and relations to a specialized sentence retrieval engine. Indexes are enriched with information extracted from linguistic analysis of the documents, including the documents from the background collection. The *merk* team applies two strategies, information retrieval (IR) and semantic web-based. For some IR-based runs they follow a two step retrieval approach. For one run they use a hypothesis generation technique. In another run they follow a majority voting scheme for selecting the correct answer from a pool of runs. In the semantic approach they use various named entity recognition techniques followed by shallow linguistic similarity, semantic similarity and network analysis approaches to identify the correct answers. The best performing strategy uses a combination of query processing followed by IR on the background collection. The *kule* team applies two different similarity-based strategies using only the input text, the question string, and the multiple choice

answers, and do not rely on any external resources or the background collection. The *nict* team experiments also with several approaches that assess the similarity of a candidate query constructed from the question plus a candidate answer to the information available in a document, or a set of background documents, to select the best answer to a multiple-choice question. They explored a range of possible similarity matching methods, ranging from simple word overlap, to dependency graph matching, to feature-based vector similarity models that incorporate lexical, syntactic and/or semantic features. The only external knowledge resource used was UMLS. The *iirg* team applies rule-based pattern matching techniques. The *lims* team adapted an existing question answering system for English named QALC, and the *ntnu* team applies TF and TF-IDF weighting schemes as well as the OMIM terms about Alzheimer for query expansion with background collections to help for machine reading comprehension. More details about the approaches applied are available in the corresponding articles in this volume.

Table 4 illustrates the mean c@1 scores for each of the 4 reading tests considering all systems. This shows the difficulty of each particular test. Test 1 at 0,11 appears to be a very hard test, whereas Test 2 at 0.34 seems to be somewhat easier.

Test 1	**Test 2**	**Test 3**	**Test 4**
0,11	0,34	0,22	0,28

Table 4. Mean c@1 scores for each reading test.

The scores per run are provided in Table 5 in terms of overall c@1, median and standard deviation of c@1, and overall accuracy.

9 Conclusions

This report presented the task *Machine Reading of Biomedical Texts about Alzheimer's Disease*, which was organised as a pilot task of the QA4MRE Lab at CLEF 2012. The task focused on biomedical texts about Alzheimer's disease in English. Participating systems should answer readability tests about the test documents provided. Each readability test consisted on 10 multiple choice questions about a document. A new corpus has been compiled for participating systems to acquire background knowledge, the Alzheimer's Disease Literature Corpus. The fact that 7 teams participated in the task with 43 runs shows that there was enough interest in machine reading exercises from research teams working on biomedical texts. The best system obtained a c@1 score of 0,55, which is certainly above baseline.

For future editions of the task we will work on improving the question design and selection of test documents. We found that some types of questions defined initially based on an example test document did not occur frequently

in the document tests used for evaluation. We will perform a deeper analysis in order to define a more stable typology of questions. Additionally, we will explore additional strategies for sampling suitable documents for the test set collections in order to select test documents that are similar in terms of contents and readability complexity. From the average scores per document we observe that some documents are more difficult than others.

Acknowledgments

This work was made possible through financial support from the University of Antwerp (GOA project BIOGRAPH). We are grateful to the organizers of the QA4MRE Lab at CLEF 2012 for hosting the pilot task. Vincent Van Asch, Florian Geitner, Cartic Ramakrishnan, Gully A.P.C. Burns, Pamela Forner, and Giovanni Moretti provided technical support. Elsevier was kind enough to allow us to include some of their articles in the background collection. We are grateful to Anita de Waard and Antony Scerri for providing the Elsevier documents. Finallly, we also thank Professor Tim Clark for providing the list of SWAN cited and derived articles.

Run	Overall c@1	Median c@1	St. Dev	Overall accuracy
Pisa12013enen	0,55	0,55	0,13	0,55
merk12062enen	0,47	0,52	0,17	0,43
merk12022enen	0,40	0,37	0,16	0,38
merk12052enen	0,39	0,47	0,27	0,35
merk12012enen	0,36	0,33	0,28	0,30
merk12072enen	0,35	0,40	0,24	0,30
kule12061enen	0,30	0,30	0,08	0,30
kule12101enen	0,30	0,30	0,08	0,30
nict12102enen	0,28	0,30	0,05	0,28
merk12042enen	0,26	0,26	0,12	0,25
iirg12021enen	0,25	0,25	0,13	0,25
iirg12041enen	0,25	0,25	0,13	0, 25
kule12041enen	0,25	0,20	0,10	0,25
kule12051enen	0,25	0,30	0,10	0,25
kule12091enen	0,25	0,20	0,19	0,25
merk12032enen	0,25	0,30	0,17	0,25
iirg12021enen	0,23	0,20	0,22	0,23
iirg12031enen	0,23	0,20	0,22	0,23
iirg12051enen	0,23	0,20	0,22	0,23
iirg12061enen	0,23	0,25	0,17	0,23
kule12031enen	0,23	0,25	0,10	0,23
nict12031enen	0,23	0,25	0,21	0,23
nict12041enen	0,23	0,25	0,17	0,23
nict12053enen	0,23	0,20	0,15	0,23
nict12063enen	0,23	0,25	0,10	0,23
nict12074enen	0,23	0,20	0,13	0,23
kule12011enen	0,21	0,19	0,12	0,18
lims12013enen	0,21	0,21	0,16	0,20
lims12024enen	0,21	0,20	0,12	0,20
lims12043enen	0,21	0,27	0,14	0,20
kule12071enen	0,20	0,20	0,08	0,20
kule12081enen	0,20	0,20	0,12	0,20
nict12091enen	0,20	0,20	0,08	0,20
ntnu12032enen	0,20	0,19	0,17	0,18
ntnu12054enen	0,20	0,20	0,16	0,20
iirg12011enen	0,18	0,10	0,15	0,18
kule12021enen	0,18	0,19	0,15	0,15
ntnu12012enen	0,18	0,15	0,17	0,18
ntnu12044enen	0,18	0,10	0,24	0,18
ntnu12022enen	0,17	0,13	0,17	0,15
nict12012enen	0,15	0,15	0,13	0,15
nict12024enen	0,15	0,15	0,13	0,15
lims12034enen	0,14	0,18	0,12	0,13

Table 5. Results per run. 'st. dev' stands for standard deviation, which is calculated over c@1 of all 4 reading tests.

Bibliography

[1] Al-Mubaid, H., Singh, R.: A new text mining approach for finding protein-to-disease associations. American Journal of Biochemistry and Biotechnology 1(3), 145–152 (2009)

[2] Barbosa-Silva, A., Fontaine, J., Donnard, E., Stussi, F., Ortega, J., Andrade-Navarro, M.: PESCADOR, a web-based tool to assist text-mining of biointeractions extracted from pubmed queries. BMC Bioinformatics 12 (2011)

[3] Bodenreider, O.: The unified medical language system (UMLS): Integrating biomedical terminology. Nucleic Acids Research 32(Suppl.1):D267D270 (2014)

[4] Etzioni, O., Banko, M., Cafarella, M.J.: Machine reading. In: Proceedings of the 21st National Conference on Artificial Intelligence. vol. 2, pp. 1517–1519. Boston, Massachusetts (2006)

[5] Gao, Y., Kinoshita, J., Wu, E., Miller, E., Lee, R., Seaborne, A., Cayzer, S., Clark, T.: SWAN: A distributed knowledge infrastructure for alzheimerdisease research. Web Semantics: Science, Services and Agents on the World Wide Web 4(3), 222–228 (2006)

[6] Peñas, A., Rodrigo, A.: A simple measure to assess the non-response. In: Proceedings of 49th Annual Meeting of the Association for Computational Linguistics - Human Language Technologies (ACL-HLT 2011). pp. 1415–1424 (2011)

[7] Peñas, A., Hovy, E.H., Forner, P., Rodrigo, Á., Sutcliffe, R.F.E., Forascu, C., Sporleder, C.: Overview of QA4MRE at CLEF 2011: Question answering for machine reading evaluation. In: CLEF (Notebook Papers/Labs/Workshop) (2011)

[8] Peñas, A., Hovy, E.H., Forner, P., Rodrigo, Á., Sutcliffe, R.F.E., Sporleder, C., Forascu, C., Benajiba, Y., Osenova, P.: Overview of QA4MRE at CLEF 2012: Question answering for machine reading evaluation. In: CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers (2012)

[9] Ramakrishnan, C., Patnia, A., Hovy, E., Burns, G.: Layout-aware text extraction from full-text pdf of scientific articles. Source code for biology and medicine 7(1), 7 (2012)

[10] Sagae, K., Tsujii, J.: Dependency parsing and domain adaptation with lr models and parser ensembles. In: Proceedings of the CoNLL 2007 Shared Task. Joint Conferences on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'07). pp. 1044–1050. Prague, Czech Republic (2007)

[11] Settles, B.: ABNER: an open source tool for automatically tagging genes, proteins and other entity names in texts. Bioinformatics 21(14), 3191–3192 (2005)

[12] Strassel, S., Adams, D., Goldberg, H., Herr, J., Keesing, R., Oblinger, D., Simpson, H., Schrag, R., Wright, J.: The DARPA machine reading program - encouraging linguistic and reasoning research with a series of reading tasks. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valletta, Malta (2010)