Reading comprehension at Entrance Exams 2015

Dominique Laurent, Baptiste Chardon, Sophie Nègre, Camille Pradel, Patrick Séguéla

Synapse Développement, 5 rue du Moulin-Bayard, 31000 Toulouse

{dlaurent, baptiste.chardon, sophie.negre, camille.pradel, patrick.seguela}@synapse-fr.com

Abstract. This article presents the participation of Synapse Développement to the CLEF 2015 Entrance Exam campaign (QA track) for English and French languages. Since fifteen years, our company works on Question Answering domain. Recently our work concentrated on Machine Reading and Natural Language understanding. Thus, the Entrance Exam evaluation is an excellent opportunity to measure the results of this work. The developed system is based on a deep syntactic and semantic analysis with anaphora resolution and some inference mechanisms. The results of this analysis are saved in sophisticated structures based on clause description (CDS = Clause Description Structure). For this evaluation, we added a dedicated module to compare CDS from texts, questions and answers. This module measures the degree of correspondence between these elements, taking into account the type of answer awaited. We participated in English and French languages. This run obtains the best results (52 good answers on 89 in English and 50 on 89 in French). So, in English and French, our system can pass the entrance exam for University!

Keywords: Question Answering, Machine Reading, Natural Language Understanding.

Introduction

The Entrance Exams evaluation campaign uses real reading comprehension texts coming from Japanese University Entrance Exams (the Entrance Exams corpus for the evaluation is delivered by NII's Todai Robot Project and NTCIR RITE). These texts are intended to be used to test the level of English of future students and represent an important part in Japanese University Entrance Exams[1]. As claimed by the organizers of the campaign: " *The challenge of "Entrance Exams" aims at evaluating systems under the same conditions humans are evaluated to enter the University"*.

We participated last year to this evaluation campaign and our results were the best for French language (33 good answers on 56) and good for English language (25 good answers on 56) but under the 50% level.

Our Machine Reading system is based on a major hypothesis: The text, in its structure and in its explicit and implied syntactic functions, contains enough information to allow Natural Language Understanding with a good accuracy. So our system does not use any external resources, i.e. Wikipedia, DbPedia and so on. Our system uses only our linguistic modules (parsing, word sense disambiguation, named entities detection and

[1] See http://www.ritsumei.ac.jp/acd/re/k-rsc/lcs/kiyou/4-5/RitsIILCS_4.5pp.97-116Peaty.pdf

resolution, anaphora resolution) and our linguistic resources (grammatical and semantic information on more than 300,000 words and phrases, global taxonomy on all these words, thesaurus, families of words, converse relation, and so on). These software modules and linguistic resources are the results of more than twenty years of development and are evaluated as the state of art for French and English.

Our Machine Reading system and the Multiple-Choice Question-Answering system needed for Entrance Exams use a database built with the results of our analysis that results in a set of Clause Description Structures (CDS) to be described in the second chapter of this article. The Entrance Exams corpus was composed this year of 19 texts with a total of 89 questions. Knowing that for each question 4 answers are proposed, the total number of choices/options was 396. Organizers of the evaluation campaign allow the systems to leave some questions unanswered if theses systems are not confident in the correctness of the answer. We did not use this opportunity but we will give in chapter 3 some results when leaving unanswered questions where the probability of the best answer is too low or lower than double of the probability of the second answer.

Machine Reading System architecture

For Entrance Exams, similar treatments are made for texts, questions and answers but the results of these treatments are saved in three different databases, allowing the final module to compare the Clause Description Structures (CDS) from text and answers to measure the probability of correspondence between CDS from text and CDS from answers. Figure 1 shows the global architecture of our system.

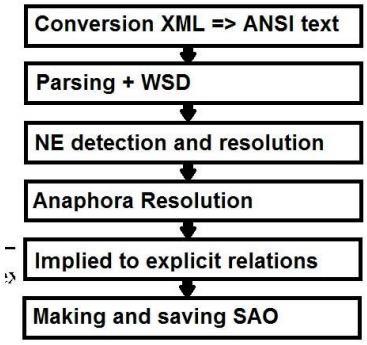

Figure 1. Description of the system

2.1 Conversion from XML into text format

The XML format allows our system to distinguish text, questions and answers. So the first operation is to extract text, then each question and the corresponding answers in text format.

2.2 Parsing, Word Sense Disambiguation, Named Entities detection

We use our internal parser which begins by a lexical disambiguation (is it a verb? a noun? a preposition? and so on) and a lemmatization. Then the parser splits the different clauses, groups the phrases, sets the part of speech and searches all grammatical functions (subject, verb, object, direct or indirect, other complements).

Then, for all polysemous words, a Word Sense Disambiguation module detects the sense of the word. For English, this detection is successful in 85 % of word senses, but for French, with a higher number of polysemous words and a higher number of senses for each word, the rate of success is about 87%). The senses disambiguated are directly linked in our internal taxonomy.

A named entity detector groups the named entities. The Named Entities detected are: names of persons, organizations and locations, but also functions (director, student, etc.), time (relative or absolute), numbers, etc. These entities are linked between them when they refer to the same entity (for example "Dominique Strauss-Kahn" or "DSK", "Toulouse" or "la Ville rose", etc). This module is not very useful for this Entrance Exams campaign except for time entities.

2.3 Anaphora resolution

We consider as anaphora all the personal pronouns (I, me, he, him, she, her, it, we, us, you, they, myself, yourself, himself, herself, itself, ourselves, yourselves, themselves), all demonstrative pronouns and adjectives (this, that, these, those), all possessive pronouns and adjectives (my, mine, his, her, its, our, ours, your, yours, their, theirs) and, of course, the relative pronouns (who, whom, whose, which, what, that) and the pronouns "one" and "ones").

During the parsing, the system builds a table with all possible referents for anaphora (proper nouns, common nouns, phrases, clauses, citations) with a lot of grammatical and semantic information (gender, number, type of named entity, category in the taxonomy, sentence where the referent is located, number of references for this referent, etc.) and, after the syntactic parsing and the word sense disambiguation, we resolve the different anaphora in the sentence by comparison with our table of referents. Our results at this step are good, equivalent or best than the state of the art. Depending of the type of pronouns and depending of the language, the rate of success of our anaphora solver is between 62 and 93%. Fortunately, for anaphora in answers, the rate of resolution has been 100% this year for English and French, due to the relative small number of referents for each anaphora.

2.4 Implied to explicit relations

When there are coordinate subjects or objects (for example "Dad and Mom"), our system keeps the trace of this coordination. For example with the coordination " Dad and Mom " the system will save three different CDS, one with the coordinate subject and two for each term of the subject. The aim of this division is to find possible answers with only one term of the coordination. But, beyond this very simple decomposition, our analyzer operates more complex operations. For example, in the sentence *" Sarah Watson was a well-known doctor who often traveled about to treat patients* ", extracted from third text

of this evaluation, our system creates four different CDs (see 2.5) using conversity and adding implied information. This mechanism exists also for the CDS structures as described in the next paragraph.

2.5 Making and saving CDS

We describe in this Section the main features of CDS structures. First we consider the attribute as an object (that could be discussed, but it allows one model of structure only). The main components of the structure are descriptions of a clause, normally compound of a subject, a verb and an object or attribute. Of course the structure allows many other components, for example indirect object, temporal context, spatial context... Each component is a sub-structure with the complete words, the lemma, the possible complements, the preposition if any, the attributes (adjectives) and so on.

For verbs, if there is some modal verb, only the last verb is considered but the modality relation is kept in the structure. Of course negation or semi-negation (forget to) are also attributes of the verb in the structure. If a passive form is encountered, the real subject becomes the subject of the CDS and the grammatical subject becomes the object. When the system encountered possessive adjective, a specific CDS is created with a link of possession. For example, in the sentence " *Sarah Watson was a well-known doctor who often traveled about to treat patients.*" (first sentence of text 17), the system will create four different CDs, the first one with "Sarah Watson" as subject, "be" as verb and "well-known doctor' on object. The second one with "Sarah Watson" as subject, "travel" as verb, "treat patients" as indirect object. The third one with "Sarah Watson" as subject, "treat" as verb, "patients" as objets and a fourth one with "doctor" as subject, "treat" as verb and "patients" as object. In fact, these relations are not very useful to find answers in this text, because the questions on this text are not medical, but it's useful at least for identify the equivalence between "Sarah Watson" and "Dr Watson" in the questions, because we have the attribute CD (Sarah Watson = doctor) and, of course, we have the equivalency Dr=doctor in our linguistic resources.

New CDS are also created when there is a converse relation. For example, in the question "*Why did the father advise his son to get a suitcase*?" (text 7), all the text is about the relation between the father (the author of the text, "I" in the text 7) and his son (the subject), so the conversity is essential to resolve anaphora and answer questions. The system manages 347 different converse relations, for example the classical "sell" and "buy", or "husband" and "wife", or "manager" and "employee", but also geographic terms (south/north, under/above...) and time terms (before/after, precedent/next...). For all these links, two CDS are created.

Links between CDS are also saved. Other relations like "cause", "judgment", "opinion" and so on are also saved and are important when the system matches the CDS of the text and the CDS of the possible answers. At the end, after all these extensions, we can consider that a real semantic role labeling is performed. This year, we spent some time to improve management of time between the sentences but this work didn't allow to really improve performance. Looking to the sentences of text 8, "*At this moment in history, science seems likely to alter our society as never before. At the same time, the power of technology has become enormous*", the system is able to identify that the time in the second sentence is the same than in the first sentence but is unable to detect the time in the first sentence, detecting only a date (hopefully not a duration), but without precision on this date (in fact here, today, actually). But in French, the system detects this date because, in place of "*At this moment in history*", the translation in French is "*Actuellement*"!

Finally the system saves also "referents", which are proper and common nouns found in the sentences, after anaphora resolution. These referents are especially useful when the

system do not find any correspondence between CDS, knowing that the frequencies in text and in usual vocabulary are arguments of the referent structures.

A specific difficulty of Entrance Exams corpus is that it is frequently spoken language with dialogs like in novels, and the number of characters is often superior to 1 or 2. For example, in the text 5, there are at least 5 essential characters : Johnny, his mother, his father, his teacher and George, which is an imaginary character !

2.6 Comparing CDS and Referents

This part of our system has been partially developed for Entrance Exams evaluation last year and this year, due to the specificity of this evaluation, specially the triple structure text/questions/answers. Once each text is analyzed, each question is analyzed, then the four possible answers are analyzed. The questions have generally no anaphora or these anaphors refer to words in the question, but the system needs to consider that "the author" (or, sometimes, "the writer") is "I" in the text. Anaphora in questions are very common and the referents are in the answer (rarely) or in the question (more commonly). For example, in the question " *Why was the family so excited about the son's trip long before it began?*", the pronoun "*it*" refers to "*trip*", and in the associated question " *Because the family considered him much too young to travel by himself.*", the pronouns "*him*" and "*himself*" refer to the son. The resolution of these anaphora is easy because the number of possible referents is low (*family, son* and *trip*), so with gender and semantic constraints, "him" can only be "*son*".

When the question is analyzed, besides the CDS structures, the system extracts the type of the question like in our Question Answering system. In Entrance Exams, these types are always non-factual types like cause ("*Why was the family so excited about the son's trip long before it began*?"), sentiment ("*What do many Japanese think of butterflies today?*"), aim ("*Why did the writer feel restless at the P.T.A. meeting?*"), signification ("*What does "youth travels light" in the last paragraph mean?*"), event ("*What happened on the yacht before the writer visited?*") and so on. Frequently, parts of the question need to be integrated into the answers. In the first sentence of text 6, for example (" *Due to the leisure brought about by technology, many people* "), the nominal group " *many people* " needs to be added at the beginning of the answers. In this case, first answer *" are puzzled by what to do with their working hours* " will become " *many people are puzzled by what to do with their working hours ”*.

Once the CDS and the type are extracted of the question, referents and temporal and spatial contexts (if they can be extracted from the question) are used to define the part of the text where elements of the answer are the most probable. For example, in the text 5 the questions 3 and 4 refer to " *the P.T.A. meeting "*, this part of speech appears only in the second half of the text, so the target of the answers is the second half, not the first one, i.e. CDS of the second half will weight more than CDS from the first half and CDS with "P.TA meeting" or only "P.T.A." will weight more.

In a first time, the system eliminates answers where there is no correspondence between CDS, referents and type of question/answer. There are few cases, only 22 on 356 answers in English and 19 in French. More generally, it seems that the method consisting in reducing the choices between answers by elimination of inadequate answers is extremely difficult to implement. Because, probably, answers are made to test the comprehension of the texts by humans and, frequently, the answer which seems to be the best choice (i.e. which integrates the bigger number of words from the text) is not the good one... and, reciprocally, the answer which seems the most distant is frequently the good one!

For the answers, two tasks are very important: adding eventually part of the question (described above) and resolution of anaphora. Hopefully the resolution of anaphora is

easiest on question and answers than in the text. The number of possible referents is reduced and, testing on the evaluation run, we found that the system made no error in French and no in English also.

To compare CDS of answers and CDS of text, we compared each CDS of text to each CDS of each answer, taking into account a coefficient of proximity of the target and the number of common elements. Subject and verb have bigger weight than object, direct or indirect, which have bigger weight than temporal and spatial context. If the system finds two elements in common, the total is multiplied by 4, if three elements are in common the total is multiplied by 16, etc. The system also increases the total when there is a correspondence with the type of the question. If only one element or no element is common to the CDS, the system takes into account the categories of our ontology, increasing the total if there is a correspondence. The total is slightly increasing if there are common referents. The total is cumulative with all the CDS of the text and finally divided by the number of CDS in the answer (often one, no more than three in the evaluation corpus).

At the end, we have, for each answer, a coefficient which ranges from 0 to 32792 (in the evaluation test, because there is no upper limit). The answer with the biggest coefficient is considered as the correct answer.

Results

Our system answered correctly to 52 questions out of 89 in English (c@1 = 0.58) and 50 questions out of 89 in French (c@1 = 0,56). The χ^2 is 33.17 in English (i.e. a probability of 0,0001 % that these results were obtained randomly) and 28,59 in French. Knowing that, randomly, a system will obtain an average 25% of good answers, in this case 22 good answers. Thus, we outperform random from 30 good answers in English and 28 in French. even if we obtain the best results in this evaluation, we don't consider these results as very good, because, excluding random, our system finds 30 good answers and do not find 37 in English, and find 28 good answers for 39 bad answers. So, for more than half of the questions, our system is unable to detect the good answer.

Considering the results files, we tested different hypothesis (see Figure 2, Results with different filters for answers). In a first hypothesis, we kept only answers where the probability of the best answer is superior or equal to 1000. In this case, we have 32 good answers on 51 questions in English. Even if the percentage of success is 63%, in fact the c@1 is equal to 0,51, which is lower than the result on 89 questions. If we keep only the questions where the probability of the best answer is superior or equal to 500, we obtain 42 good answers on 68. In this case, results are not better: the percentage of success is 62% but the c@1 is equal to 0,58, equivalent to our result of 0,58 on the total of questions. Finally we kept only the answers where the probability of the best answer is almost twice the probability of the second best answer. In this case, we obtained 22 good answers on 31, which is a good result with 71% of successful answers but a c@1is equal to 0.41 because the system answers only third of the questions!

		Results	% successful	c@1
English	evaluation run	52/89	58 %	0.58
	probability >= 1000	32/51	63 %	0.51
	probability >= 500	42/68	62 %	0.58
	best >= double 2nd	22/31	71 %	0.41
French	evaluation run	50/89	56 %	0.56

	probability >= 1000	27/41	66 %	0.47
	probability >= 500	36/59	61 %	0.54
	best >= double 2nd	21/30	70 %	0.39

Figure 2. Results with different filters for answers.

Finally, keeping all the questions and all the answers was the best strategy and our system passes the Entrance Exams for the Japanese University in both languages! If we closely look at the results on a text by text basis , 17/19 are superior or equal to 50% in English and 16/19 in French. For English, the system finds all the answers for text 14 but finds only 1 answer on text 1 (25%) and 2 answers on text 11 (33%). For French, the system finds all the answers for the text 16 but finds only 1 answer for text 1 (25%) and for text 12 (25%) and 2 out of five for text 19 (40%).

If our system obtains a score sufficient to pass the Entrance Exam, there is an area where the computer is clearly superior to the human: speed. The French run is executed in 4,8 seconds, which means a speed of about 2500 words by second and the English one in 3.9 seconds (about 3000 words by second). Because we did not try to optimize the code, this speed could be better (the speed of our parser is more than 10 000 words by second), specially if we rewrite the comparison between CDS of text and CDS of answers.

Analysis of results

Strangely, this year, like in 2014 and 2013, there were 5 participants, resulting in 19 runs this year, versus 29 runs in 2014 and 10 runs in 2013. In 2013, on these 10 runs, 3 obtain results superior to random and 7 inferior or equal to random. In 2014, out of 29 runs, 14 obtain results superior to random and 15 inferior or equal to random. This year, 16 runs obtain results superior to random and only 3 inferior to random. Comparing our results from last year, we think that the difficulty of the test is similar to the difficulty of last year, so we can notice a significant improvement of the global results this year.

Anyway the difficulty of the task is considerable. These tests have been written by humans to evaluate the reading comprehension of humans. So, for example, the answer which seems the best, i.e. which includes the higher number of words from the text, is generally a bad answer. And even if we obtained the best results, we have seen that, excluding random, our system fails more often than he successes !

To demonstrate that with our run, we will take two examples, the first one is very basic, the second one is more complex. As you can imagine, our system finds the good answer in the first case in English and French, not in the second case. The easiest question/answer is the first question of the first text:

The woman telling the story

1 *always went shopping with her family on Fridays.*
2 *had been very busy and needed some time to recover.*
3 *wanted a newspaper and some chocolate to take home to her family.*

4. bought a newspaper and some chocolate so that she could keep a place at the table.

The reference part of text is :

"*Last Friday, after doing all the family shopping in town, I wanted a rest before catching the train, so I bought a newspaper and some chocolate and went into the station coffee shop -that cheap, self-service place with long tables to sit at. I put my heavy bag down on the floor, put the newspaper and chocolate on the table to keep a place...*"

The main difficulty for this question is the formulation of the question "the woman telling the story". Our system is unable to identify this part of speech with the author of the text (I in the text). Fortunately the text has only two characters : a woman (identified by "my husband") and a young man. Probably the results were different if the two characters were women ! The probabilities of the answers are, in English, 678 for answer 1 (due to 2 CDs and 3 referents : shopping, family, Friday), 84 for answer 2 (only one CD and one referent : rest, synonym of "recover"), 1239 for answer 3 (two CDs and three referents : newspaper, chocolate, family), 2804 for answer 4 (three CDs and six referents : buy, newspaper, chocolate, keep, place, table). Even if we didn't take into account CDs and having a very simple approach like "bag of words", the answer 4 is the more probable. And our system works fine, In English like in French, for similar simple questions !

The second example is considerably more complex and our system didn't find the good answer. It considers the first question for the text 13:

What had Mark Wellman long desired to do?

1 *To accomplish one of the most difficult rock climbs in the world.*
2 *To be the first to conquer El Capitan.*
3 *To climb the highest mountain in California.*
4 *To help his friend Peter climb El Capitan.*

The reference text is :

Peter Corbett helped Mark Wellman out of his wheelchair and onto the ground. They stood before El Capitan, a huge mass of rock almost three-quarters of a mile high in California's beautiful Yosemite Valley. It had been Mark's dream to climb El Capitan for as long as he could remember. But how could a person without the use of his legs hope to try to climb the highest vertical cliff on earth?

The good answer is 1 and the system in English returns 4 and in French returns 3.Why ? The first problem comes from the type of the question. If, in French, the system identifies the question as "aim", in English the type of the question is identified as event... Secondly there are CDs corresponding to all the possible answers. The only one which can be excluded is the answer 4 because, in the text, "Peter helps Mark" and never "Mark helps Peter" but, in English, the association of the question with answer returns "what had Mark Wellman long desired to do to help his friend Peter climb El Capitan" and the sequence "desired to do to help" is too complex for CDs, so this answer is not excluded. Notice finally that the correspondence between "*one of the most difficult rock climbs in the world"* and *" the highest vertical cliff on earth*" is not evident, even with good thesauri !

Errors and translations

Last year, our results for French were clearly better than for English. This year, the results are similar (52 good answers versus 50 good answers). Even if, since last year, we worked a lot on English, improving the linguistic resources, the parsing and several components, it seemed to us interesting to compare the answers to the questions between French and English.

The Figure 3 below lists the results compared by language:

	Good answers
In English and in French	42
only in English	10
only in French	8
nor in English or French	29

Figure 3. Good answers by language

Globally the results in English are closed to results obtained in French. That seems logical, knowing that we use similar methods and structures for the two languages. For the questions where results are different, sometimes it depends of very little differences in cumulative scores. For example, for the question 1 of text 17, the global score in English for answer 2 is 512 and the global score for answer 3 is 543. In French the global score for answer 2 is 924 and the global score for answer 3 is 810. So in English the answer is incorrect and in French it is correct. Looking into details at the scoring, we found out that the difference comes from the formulation of the answer 2 in English ("*She wanted to travel by herself.*") and the formulation in French ("*Elle voulait voyager seule.*"). In French, we have in the text "*elle serait heureuse d'y rester seule si elle le pouvait.*" and in English "s*he would be glad to be alone if she could*." The similarity in French between the word "*seule*" in the text and the question, which didn't exist in English justify the difference of scoring.

Looking at the questions where the results in English are correct and incorrect in French, we discovered that, frequently, this difference is due to the quality of the translation. For example, looking to the question 1 of text 8 ("*The advantage of using earphones on a train is that* :"), the good answer found in English is the answer 4 "*you seldom bother people around you*" and in French it's the answer 2 *("l'on peut avoir un meilleur contact avec son appareil."*). The answer 4 in French is translated as "*l'on ne dérange rarement les gens autour de soi*." which seems a good translation, but is interpreted by the semantic parser as the inverse of English, because "ne" and "rarement" are interpreted as a double negation, equivalent to "*on dérange les gens autour de soi*" in place of "*on dérange rarement les gens autour de soi*".

Conclusion

All the software modules and linguistic resources used in this evaluation exist since many years and are the property of the company Synapse Développement. The parts developed for this evaluation are the Machine Reading infrastructure, some improvements of the resolution of anaphora in English and the complete module to compare CDS from text and answers. No external resources have been used.

With 52 good answers on 89 English questions and 50 good answers for 89 French questions, the results are good. However, the limitations of the method appears clearly, knowing that random gives 25 % good answers.

Acknowledgements. We acknowledge the support of the CHIST-ERA project "READERS Evaluation And Development of Reading Systems" (2012-2016) funded by ANR in France (ANR-12-CHRI-0004) and realized with the collaboration of Universidad del Pais Vasco, Universidad Nacional de Educación a Distancia in Madrid, and University of Edinburgh. This work benefited from numerous exchanges and discussions with these partners led within the framework of the project.

References

1 Arthur, P., Neubig, G., Sakti, S., Toda, T., Nakamura, S., NAIST at the CLEF 2013 QA4MRE Pilot Task. *CLEF 2013 Evaluation Labs and Workshop Online Working Notes, ISBN 978-88-904810-5-5,* ISSN 2038-4963, Valencia -Spain, 23 -26 September, 2013 (2013)

2 Banerjee, S., Bhaskar, P., Pakray, P., Bandyopadhyay, S., Gelbukh, A., Multiple Choice Question (MCQ) Answering System for Entrance Examination, Question Answering System for QA4MRE@CLEF 2013. *CLEF 2013 Evaluation Labs and Workshop Online Working Notes, ISBN 978-88-904810-5-5, ISSN 2038-4963*, Valencia -Spain, 23 -26 September, 2013 (2013)

3 Berant J., Srikumar V., Chen P-C., Vander Linden A., Harding B., Huang B., Clark P., Manning C.D., A., Modeling Biological Processes for Reading Comprehension, *EMNLP 2014.*

4 Buck, G., Testing Listening Comprehension in Japanese University Entrance Examinations, *JALT Journal, Vol. 10, Nos. 1 & 2*, 1988 (1988)

5 Iftene, A., Moruz, A., Ignat, E.: Using Anaphora resolution in a Question Answering system for Machine Reading Evaluation. *Notebook Paper for the CLEF 2013 LABs Workshop* -QA4MRE, 23-26 September, Valencia, Spain (2013)

6 Indiana University, French Grammar and Reading Comprehension Test. *http://www.indiana.edu/~best/bweb3/french-grammar-and-reading-comprehension-test/*

7 Kobayashi, M., An Investigation of method effects on reading comprehension test performance, *The Interface Between Interlanguage, Pragmatics and Assessment: Proceedings of the 3rd Annual JALT Pan-SIG Conference*. May 22-23, 2004. Tokyo, Japan: Tokyo Keizai University (2004)

8 Laurent, D., Séguéla, P., Nègre, S., Cross Lingual Question Answering using QRISTAL for CLEF 2005 *Working Notes, CLEF Cross-Language Evaluation Forum, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, 20-22 september 2006, Alicante, Spain (2006)

9 Laurent, D., Séguéla, P., Nègre, S., Cross Lingual Question Answering using QRISTAL for CLEF 2006, Evaluation of Multilingual and Multi-Modal Information Retrieval Lecture Notes in Computer Science, *Springer, Volume 4730*, 2007, pp 339-350 (2007)

10 Laurent, D., Séguéla, P., Nègre, S., Cross Lingual Question Answering using QRISTAL for CLEF 2007, *Working Notes, CLEF Cross-Language Evaluation Forum, 7th Workshop of the Cross-Language Evaluation Forum*, CLEF 2008, Budapest, Hungary (2008)

11 Laurent, D., Chardon B., Nègre S., Séguéla, P., French run of Synapse Développement at Entrance Exams 2014, *Working Notes for CLEF 2014 Conference, CEUR Workshop Proceedings, Vol. 1180,* Sheffield, UK, September 15-18, 2014, pp. 1415-1426

12 Laurent, D., Chardon B., Nègre S., Séguéla, P., English run of Synapse Développement at Entrance Exams 2014, *Working Notes for CLEF 2014 Conference, CEUR Workshop Proceedings, Vol. 1180,* Sheffield, UK, September 15-18, 2014, pp. 1404-1414

13 Li, X., Ran, T., Nguyen, N.L.T., Miyao, Y., Aizawa, A., Question Answering System for Entrance Exams in QA4MRE. *CLEF 2013 Evaluation Labs and Workshop Online Working Notes, ISBN 978-88-904810-5-5,* ISSN 2038-4963, Valencia -Spain, 23 -26 September, 2013 (2013)

14 MacCartney, B., Natural Language Inference, PhD Thesis, Stanford University, June 2009 (2009)

15 Mulvey, B., A Myth of Influence: Japanese University Entrance Exams and

Their Effect on Junior and Senior High School Reading Pedagogy, JALT Journal, Vol. 21, 1, 1999 (1999)

16 National Institute of Informatics, Todai Robot Project, NII Today, n°46, July 2013 (2013)

17 Peñas, A., Hovy, E., Forner, P., Rodrigo, Á., Sutcliffe, R., Sporleder, C., Forascu, C., Benajiba, Y., Osenova, P.: Overview of QA4MRE at CLEF 2012: Question Answering for Machine Reading Evaluation. CLEF 2012 *Evaluation Labs and Workshop Working Notes Papers*, 17-20 September, 2012, Rome, Italy (2012)

18 Peñas, A., Miyao, Y., Hovy, E., Forner, P., Kando, N. : Overview of QA4MRE at CLEF 2013 Entrance Exams Task. *CLEF 2013 Evaluation Labs and Workshop. Online Working Notes, ISBN 978-88-904810-5-5*. ISSN 2038-4963 (2013)

19 Peñas, A., Hovy, E., Forner, P., Rodrigo, Á., Sutcliffe, R., Sporleder, C., Forascu, C., Benajiba, Y., Osenova, P.: Evaluating Machine Reading Systems through Comprehension Tests. *LREC 2012 Proceedings of the Eight International Conference on Language Resources and Evaluation*, 21-27 May, 2012, Istambul (2012)

20 Peñas, A., Rodrigo, Á. : A Simple Measure to Assess Non-response. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 1415–1424*, Portland, Oregon, June 19-24, 2011. Association for Computational Linguistics (2011)

21 Quintard, L., Galibert, O., Adda, G., Grau, B., Laurent, D., Moriceau, V., Rosset, S., Tannier, X., Vilant, A. , Question Answering on Web Data : The QA Evaluation in Quero, *Proceedings of the Seventh Conference on Language Resources and Evaluation*, 17-23 May, 2010, Valletta, Malta (2010)

22 Riloff, E., Thelen, M., A Rule-based Question Answering System for Reading Comprehension Tests. Proceedings of ANL P/NAACL 2000. Workshop on Reading Comprehension Tests as Evaluation for computer-Based Language Understanding Systems, PP. 13-19 (2000)