IDRAAQ: New Arabic Question Answering system based on Query Expansion and Passage Retrieval

Lahsen Abouenour[1], Karim Bouzoubaa[1] and Paolo Rosso[2]

[1] Mohammadia School of Engineers, Med Vth University-Agdal, Rabat, Morocco
[2] Natural Language Engineering Lab., ELiRF, Universitat Politècnica de València, Spain
abouenour@yahoo.fr, karim.bouzoubaa@emi.ac.ma, prosso@dsic.upv.es

Abstract. Arabic is one of the languages which are less concerned by researchers in the field of Question Answering. The paper presents core modules of a new Arabic Question Answering system called IDRAAQ. These modules aim at enhancing the quality of retrieved passages with respect to a given question. Experiments have been conducted in the framework of the main task of QA4MRE@CLEF 2012 that includes this year the Arabic language. Two runs were submitted. Both runs only use reading test documents to answer questions. The difference between the two runs exists in the answer validation process which is more relaxed in the second run. The Passage Retrieval (PR) module of our system presents multi-levels of processing in order to improve the quality of returned passage and thereafter the performances of the whole system. The PR module of IDRAAQ is based on keyword-based and structure-based levels that respectively consist in: (i) a Query Expansion (QE) process relying on Arabic WordNet semantic relations; (ii) a Distance Density N-gram Model based passage retrieval system. The latter level uses passages retrieved on the basis of QE queries and re-ranks them according to a structure-based similarity score. Named Entities are recognized by means of a mapping between the YAGO ontology and Arabic WordNet. The experiments that we conducted show that with respect to the accuracy and c@1 measure, IDRAAQ registered encouraging performances in particular with factoid questions. The same experiments allowed us to identify the lacks of the system especially when processing non factoid questions and at the Answer Validation stage. The IDRAAQ system, which is still under construction, will integrate a Conceptual Graph-based passage re-ranking introducing a semantic level to its PR module.

Keywords. Arabic Question Answering, Passage Retrieval, Query Expansion, Distance N-gram Density Model, Arabic WordNet.

1 Introduction

Question Answering for Machine Reading Evaluation (QA4MRE) at CLEF 2012 is the fourth campaign which represents an evolution of previous evaluation approaches in Natural Language Processing (NLP), including Question Answering, Recognizing-Textual Entailment and Answer Validation. Like previous editions, the campaign provides large document collections that serve as a background for each particular reading test. Indeed, Machine Reading requires a deeper analysis and inference of text and in turn may need background knowledge acquisition.

The 2012 test set is composed of 4 topics, namely "Aids", "Climate change" and "Music and Society" -the same topics adopted last year- plus one additional new topic, namely "Alzheimer". This year is also particular in that two languages have been added: Arabic and Bulgarian in addition to the previously considered languages namely English, German, Italian, Romanian and Spanish. Materials are exactly the same in all languages, created using parallel translations.

We have participated in the main task in order to evaluate an ongoing Arabic QA system called IDRAAQ: Information and Data Reasoning for Answering Arabic Questions. As it is an under construction project, only two runs have been submitted. The two runs have not considered any background collection. Answers were searched within the documents of the reading test in concern.

Section 2 presents an overview of IDRAAQ. Section 3 describes the main tools and resources used in this system. The experiments carried out on test data sets are discussed in Section 4 along with the results. The conclusions are drawn in Section 5.

2 Overview of the IDRAAQ system

2.1 System Architecture

The IDRAAQ[1] system is fully programmed in Java. The system also makes use of other third party components and resources. The system is designed around the three typical modules of a Question Answering system, namely (see Figure 1):

(i) Question analysis and classification module. In this module a question is analyzed in order to extract its keywords, identify the structure of the expected answer and form the query to be passed to the PR module.

(ii) Passage Retrieval (PR) module. This module is one of the most important components of a Q/A system. The quality of the results returned by such system depends mainly on the quality of the PR module. Indeed, this module uses the query formed by the previous module and extracts a list of passages from an Information Retrieval process (generally a Search Engine such as Google[2] or Yahoo![3]). Thereafter,

[1] The word "IDRAAQ" in Arabic has the following meanings and senses: to understand, to recognize, to reach an objective, knowledge, intelligence, etc.

[2] http://www.google.com

[3] http://www.yahoo.com

this module has to perform a ranking process in order to improve the relevance of the candidate passages according to the user question.

(iii) Answer Validation (AV) module. This module tries to validate an answer from a list of candidate answers relying on passages that are provided by the previous module.

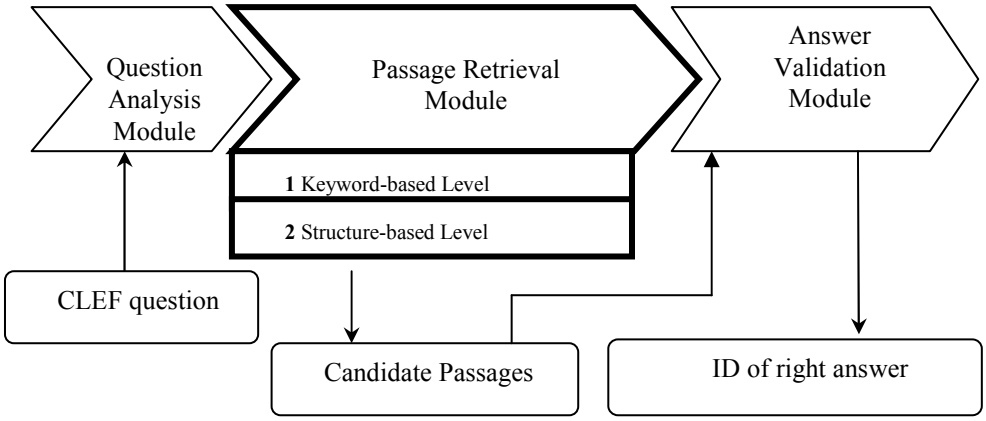

Fig. 1. The three modules of the IDRAAQ system

Since the PR module provides candidate passages in which the Answer Validation module tries searching the right answer, the performance of the IDRAAQ system is mainly dependant on this module and on the quality of its returned passages. As illustrated in Figure 1, the PR module of IDRAAQ is formed by two implemented levels: keyword-based level (Label 1) and structure-based level (Label 2). The former integrates a semantic QE process and the latter uses a Distance Density N-gram based PR tool.

Another level (third level) is under construction within the IDRAAQ system: the semantic reasoning level. It is based on comparing representations of question and candidate passages in terms of Conceptual Graphs (CGs) (Sowa, 1984) through projection and generalization operations. Since this level is on its building and testing stage, we did not consider the corresponding process in the current edition of QA4MRE. Therefore, in the following sub sections we only provide details about the first two levels.

2.2 Keyword-based level

This level is concerned with a semantic Query Expansion (QE) process. Each question keyword is substituted by its semantically related terms that are extracted from the Arabic WordNet (AWN) (Elkateb et al., 2006). In AWN, four relations are used in this level: synonymy, hyponymy, hypernymy and SUMO-AWN relations. SUMO

(Suggested Upper Merged Ontology) is a high level ontology mapped with AWN synsets[4]. Figure 2 is an illustration of the objective of our QE process.

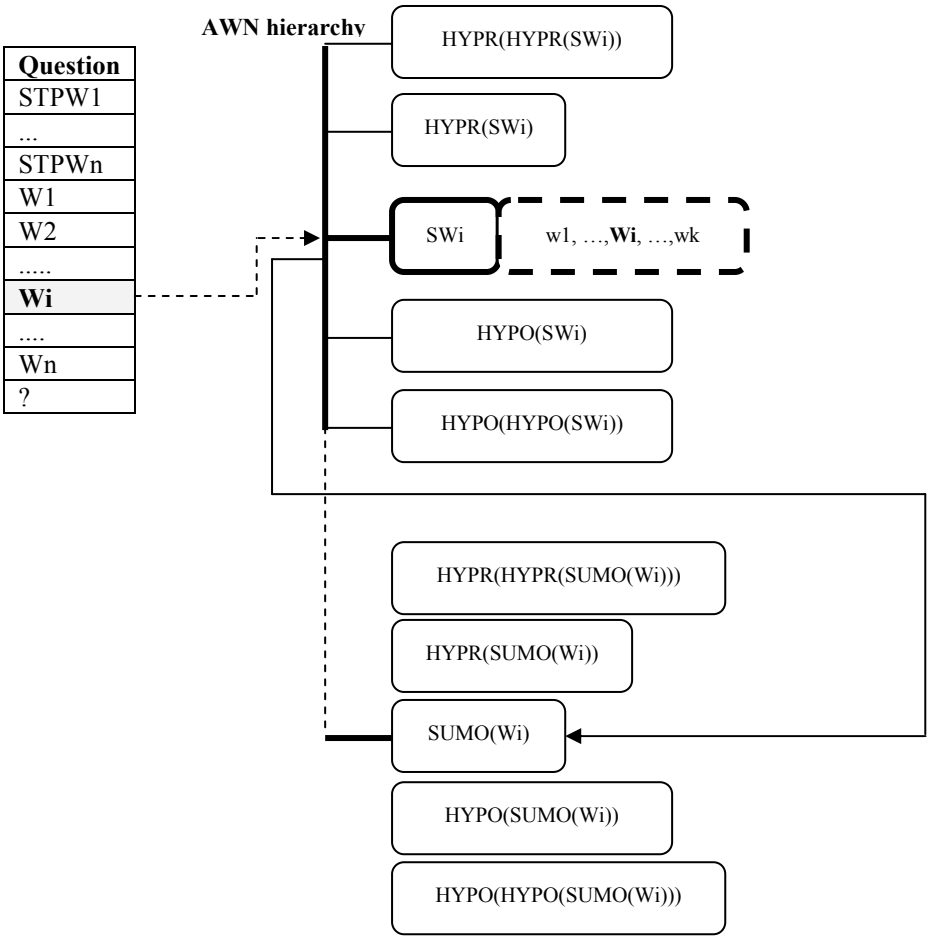

Fig. 2. AWN-based QE integrated in the IDRAAQ system

As illustrated in Figure 2, from each question, we only consider non stopwords (STPWn) in the QE process. Concretely, the AWN-based QE process accepts as input an Arabic word (non stop words extracted from the question), say Wi and generates the following terms:

a. Morphological variants of Wi using "AL KHALIL" system[5];
b. Words that share the same AWN synsets (SWi) with Wi (the synonyms w1,. ..wk);

[4] In AWN a synset is a group of synonyms that can be used in a specific context. Each word can have many senses according to the synset to which it belongs.

[5] http://sourceforge.net/projects/alkhalil/

c. Words that share the AWN synsets that are hyponyms of each SW_i; Let us refer to these synsets by HYPO(SW_i);
d. Words that share the AWN synsets that are hypernyms of each SW_i; These synsets are referred to by HYPR(SW_i);
e. Words that appear in the definition of the SUMO concept which is equivalent to each SW_i.

The same process is again performed for words related to HYPO(SW_i) and HYPR(SW_i). Note that in order to avoid endless recursive process we move just 2 levels up and down in the AWN hierarchy starting from the synset SW_i. In this way, for each question keyword, we generate a list of words that represent the context of the keyword in the AWN hierarchy as well as semantically related terms in other similar contexts in this hierarchy.

This process extracts the words belonging to the context of the expanded word by moving up and down in the AWN hierarchy. In order to catch other contexts that are semantically related to the context of the original word (i.e., W_i), we rely on the SUMO concept (SUMO(W_i)) which is linked to SW_i. In SUMO, each concept has a definition which involves many other SUMO concepts. By moving to the synsets that are equivalent to these latter concepts, we can get other semantically related words.

The semantic QE process illustrated in Figure 1 results in a number of new terms. These terms are used to form new queries by substituting a keyword in the question by its related terms. Note that in the case of Named Entities (NEs) keywords, we substitute the keyword just by its synonyms. The hypernyms are just added before the keyword in the question. This is due to the fact that a hypernym of a NE is usually its category (for instance person, country, etc.).

IDRAAQ uses an enriched version of AWN. This enrichment mainly concerns NEs, noun hyponymy relations and verbs. As factoid questions represent high percentage of processed questions, a mapping between AWN and the large English NE ontology called YAGO[6] was done and was part of the considered AWN release.

2.3 Structure-based level

The objective of this level is filtering the passages that would be returned after applying level 1. As mentioned above, for each question, different new queries are generated according to the terms extracted from AWN. These queries are important in number but are not all relevant for the question that may lead in considering irrelevant passages. Thus, the structure-based level introduces a new criterion to efficiently rerank passages: the Distance N-gram Density (Gomez et al., 2005). This model considers sequence of n adjacent words (n-gram) extracted from a sentence or a question. All possible n-grams of the question are searched. It also assigns them a score according to the n-grams and weight that appear in the retrieved passages.

[6] Yet Another Great Ontology: available at http://www.mpi-inf.mpg.de/YAGO-naga/YAGO/downloads.html

If a passage contains one or more related terms (those generated by the AWN-based QE process) then it is retrieved. However, the relevancy of this passage depends on the structure in which these terms appear. The more this structure is similar to the one of the question, the more relevant the passage is considered.

In the IDRAAQ system, this model is implemented through the Java Information Retrieval System (JIRS) (Gomez et al., 2005). This language independent system underwent some adaptations in order to be used in the context of the Arabic language (Benajiba et al. 2007). The main modifications were made on the Arabic language-related files (text encoding, stop-words, list of characters for text normalization, Arabic special characters, question words, etc.).

The JIRS is integrated in the PR module of IDRAAQ following many steps:

- Step 1: extract related queries of a question;
- Step 2: the list of queries is formatted using the JIRS input file;
- Step 3: documents are also formatted using the SGML JIRS format so that a collection of documents is built;
- Step 4: the collection built in step 3 is indexed using the corresponding JIRS process;
- Step 5: the JIRS "PassageSearch" process is performed on the indexed collection and using the input file. We customize the system to only the first five passages are retrieved for each query in the input file;
- Step 6: over all the queries, the five passages, with the best JIRS similarity score, are considered in the Answer Validation module.

3 Evaluation

The 2012 test set is composed of 4 topics; each topic includes 4 reading tests. Each reading test consists of one document, accompanied by 10 questions, each with a set of five answer options per question. Therefore, for each language task, there are in total:

- 16 test documents (4 documents for each of the four topics)
- 160 questions (10 questions for each document)
- 800 choices/options (5 for each question)

Questions have the following characteristics:

- They are in the form of multiple choice, where for each question, 5 possible answers are given;
- They are designed so that focus on testing the comprehension of one single document;
- Test the reasoning capabilities of systems, which means that inferences, relative clauses, elliptic expressions, meronymy, metonymy, temporal and spatial reasoning, and reasoning on quantities may be exploited;

- They may involve background knowledge, i.e., information that is not present in the test document given. In such cases, information from the Background collections is needed to fill in the knowledge gap to answer the question.

Questions may be of the following types:

1. **FACTOID**: Where or When or By--Whom
2. **CAUSAL**: What was the cause/result of Event X?
3. **METHOD**: How did X do Y? Or: In what way did X come about?
4. **PURPOSE**: Why was X brought about? Or: What was the reason for doing X?
5. **WHICH IS TRUE**: Here one must select the correct alternative from a number of statements, e.g. What can a 14 year old girl do?

The IDRAAQ system applies for each question the preprocessing stage, the keyword-based stage and the structure-based stage. The answer checking process matches candidate answers with returned passages. The first run that we have submitted uses a strict answer checking process while the second introduces a relaxation especially when the answer is composed of more than two words.

Each test receives an evaluation score between 0 and 1 using c@1 (Peñas et al., 2011). This measure, already tried in previous CLEF QA Tracks, encourages systems to reduce the number of incorrect answers while maintaining the number of correct ones by leaving some questions unanswered. Systems receive evaluation scores from two different perspectives:

1. At the question-answering level: correct answers are counted individually without grouping them;
2. At the reading-test level: figures both for each reading test as a whole and for each separate topic are given.

Thus, two measures have been considered as follows:

- Overall Accuracy which is calculated using the formula:

  Accuracy = nr/n

where:
nr: is the number of correctly answered questions
n: is the total number of questions

- The c@1 measure which is represented by the formula:

  C@1 = (nr + nu * (nr/n)) / n

where:
nu: is the number of unanswered questions

Obtained results also presents number of unanswered question with right and wrong candidate answers. However, in both runs, we did not consider this possibility in the submitted outputs.

Table 1 and 2 presents the obtained results in terms of: (i) accuracy over all questions and (ii) the overall as well as detailed c@1 measure.

RUNS	OVERALL ACCURACY	ANSWERED		UNANSWERED		
		RIGHT	WRONG	EMPTY	RIGHT	WRONG
run #1	0.08	12	21	127	0	0
run #2	0.13	21	49	90	0	0

Table 1. Overall accuracy of IDRAAQ over the two submitted runs

RUNS	c@1 measure				
	Overall	Topic #1	Topic #2	Topic #3	Topic #4
run #1	0.13	0.25	0.18	0.05	0.05
run #2	0.21	0.36	0.19	0.08	0.17

Table 2. Overall and detailed c@1 related to IDRAAQ

As shown in Table 1 above, the overall accuracy reaches 0.13 in the second run. This accuracy is calculated over the 160 questions. If we only consider the 70 answered questions (21+49 in Table 1), the accuracy is 0.30 in the case of run #2.

Regarding the c@1 measure, Table 2 shows the overall of 0.21 as of the second run (versus 0.13 for the first run). With respect to this measure, our system registered different performances over the four topics. Indeed, from Table 2 the maximum value was reached over Topic #1 (i.e. AIDS) in the two runs (0.25 in run #1 versus 0.36 in run #2).

At reading-test level, our system reached its best value of c@1 measure when answering questions belonging to topic #1 (i.e., AIDS). Figure 3 illustrates a comparison between the best c@1 measures obtained over the four topics with respect to this level. Topic #3 is the one for which lower performances have been reached.

Let us analyze questions for which our system succeeds and those for which it fails, i.e., questions belonging to the above topics (i.e., topic #1 and #3).

From this analysis, most of the answered questions are factoid ones (When, Who, What, etc.). This shows that using Arabic WordNet mapped with YAGO (which con-

tains high number of Named Entities) has a positive impact on system performances especially when processing factoid questions.

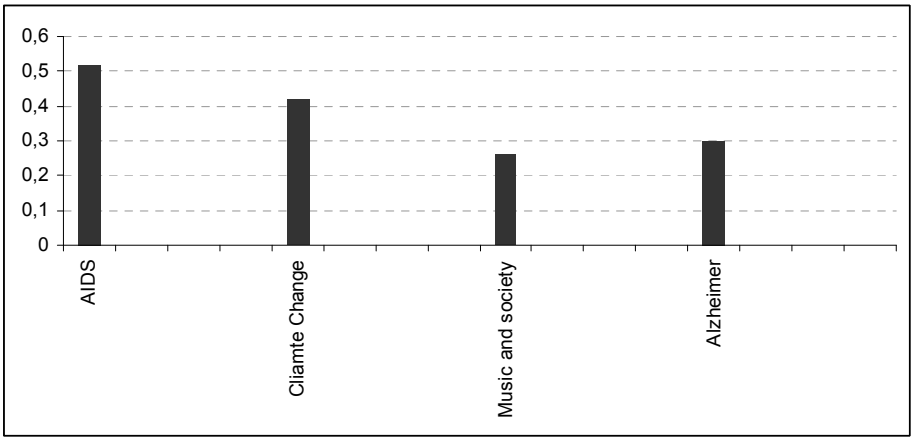

Fig. 3. Best c@1 obtained in reading tests over topics

On the other hand, the questions where the system fails to get a correct answer falls into five categories:

- Questions that are not factoid such as LIST questions (questions starting with Give a list of ...) and REASON questions (questions starting with Why ...);
- Questions with translation errors. For instance, in reading-test #4 question #4 the translation of "What is the mechanism by which HIV-positive Brazilians receive free ARV drugs?" is " ماهي الأليات المستعملة لإعطاء البرازيليين المصابين بداء نقصان المناعة البشرية المضادة للفيروسات القهقرية مجاناً؟" which is not an understandable Arabic question. This remark can also be applied on reading-test documents.
- Questions not starting with question stopword (such as What, When, etc.). For example, reading-test #6 question #3 " ووفقا للحكومة البرازيلية، ما هي الأسباب الرئيسية لتغير المناخ؟" (According to the Brazilian government, what is one the main reasons for climate change?)
- Questions with long candidate answers. For instance, questions #3 and #4 in reading-test #13 " ما هو النظام الغذائي الذي يمكن أن يخفض من خطر الإصابة بمرض الزهايمر؟" (What type of diet may reduce the risk of Alzheimer's disease?) and "لماذا لا يوصى باستعمال أنابيب التغذية لمرضى الزهايمر الذين لديهم صعوبات في البلع؟" (Why are feeding tubes not always recommended for Alzheimer's patients who have difficulties with swallowing?).

4 Conclusion

The current edition of QA4MRE has considered for the first time the Arabic language. We took advantage from this opportunity to test our semantic QE process combined with the Distance N-gram Density model. The obtained results are encour-

aging in particular for factoid questions. The analysis of IDRAAQ system performances allowed us to identify the category of questions in which the system fails to validate the right answer.

According to previous preliminary experiments (Abouenour et al., 2009), the integration of the third level based on Conceptual Graphs and semantic similarity would improve the performances of the system at the PR module as well as the Answer Validation module. Indeed, representing knowledge in the question and candidate passages would help in comparing them at a semantic level which is more advanced than the keyword and structure levels that we have considered in this experiment. The CLEF 2012 Gold standard for the Arabic language will help us in pre-testing the capabilities of the system with the third level as well as the use of background collection and other resources for answering the questions.

The perspective of the current work is preparing the system in order to participate in the next edition of QA4MRE for Arabic in an aim of reaching maturity of the best well-known QA systems for other languages.

Acknowledgment

The European Commission as part of the WIQ-EI IRSES-Project (grant no. 269180) within the FP 7 Marie Curie People Framework has partially funded the work of the third author. His work was carried out also in the framework of the MICINN Text-Enterprise (TIN2009-13391-C04-03) research project and the Microcluster VLC/Campus (International Campus of Excellence) on Multimodal Intelligent Systems.

References

Abouenour L., Bouzoubaa K. and Rosso P. (2009a). *Three-level approach for Passage Retrieval in Arabic Question /Answering Systems*. In Proc. of the 3rd International Conference on Arabic Language Processing CITALA2009, Rabat, Morocco.

Abouenour L., Bouzoubaa K. and Rosso, P. (2009b). *Structure-based evaluation of an Arabic semantic Query Expansion using the JIRS Passage Retrieval system*. In Proc. Workshop on Computational Approaches to Semitic Languages, E-ACL-2009, Athens, Greece, April, 2009. Published by the Association for Computational Linguistics (ACL), pp. 62-68, Stroudsburg, PA, USA.

Abouenour L., Bouzoubaa K. and Rosso, P. (2010a). *An evaluated semantic QE and structure-based approach for enhancing Arabic Q/A*. In the Special Issue on "Advances in Arabic Language Processing" for the IEEE International Journal on

Information and Communication Technologies (IJICT), ISSN: 0973-5836, Serial Publications.

Abouenour L., Bouzoubaa K. and Rosso, P. (2010b). *Using the Yago ontology as a resource for the enrichment of Named Entities in Arabic WordNet*. Workshop on Language Resources (LRs) and Human Language Technologies (HLT) for Semitic Languages Status, Updates, and Prospects, LREC'10 Conference, Malta.

Abouenour, L. (2011). *On the Improvement of Passage Retrieval in Arabic Question/Answering (Q/A) Systems*. Lecture Notes in Computer Science, 2011, Volume 6716/2011, 336-341, DOI: 10.1007/978-3-642-22327-3_50. R. Muñoz et al. (Eds.), NLDB'11. Springer-Verlag, Berlin-Heidelberg.

Benajiba Y., Rosso P. and Gómez, J.M. (2007). *Adapting JIRS Passage Retrieval System to the Arabic*. In Proc. 8th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2007, Springer-Verlag, LNCS(4394), pp. 530-541.

Elkateb, S., Black, W., Vossen, P., Farwell, D., Rodríguez, H., Pease, A., Alkhalifa, M. (2006). Arabic WordNet and the Challenges of Arabic. In proceedings of Arabic NLP/MT Conference, London, U.K.

Gomez J. M., Montes-Gomez M., Sanchis E., Villasenor-Pineda L. and Rosso P. (2005). *Language independent passage retrieval for question answering*. In Fourth Mexican International Conference on Artificial IntelligenceMICAI 2005, Lecture Notes in Computer Science, pages 816–823, Monterrey, Mexico, 2005. Springer Verlag.

Peñas, A. and Rodrigo, A. A Simple Measure to Assess Non-response. In Proceedings of 49th Annual Meeting of the Association for Computational Linguistics-Human Language Technologies (ACL-HLT 2011), Portland, Oregon, USA, June 19-24, 2011.

Sowa John F. (1984). Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley Company.