

An Approach to Answer Selection in Question-Answering Based on Semantic Relations

Ana Cristina Mendes and Luísa Coheur

Spoken Language Systems Laboratory - L²F/INESC-ID
 Instituto Superior Técnico, Technical University of Lisbon
 R. Alves Redol, 9 - 2º - 1000-029 Lisboa, Portugal
 ana.mendes@l2f.inesc-id.pt – luísa.coheur@l2f.inesc-id.pt

Abstract

A usual strategy to select the final answer in factoid Question-Answering (QA) relies on redundancy. A score is given to each candidate answer as a function of its frequency of occurrence, and the final answer is selected from the set of candidates sorted in decreasing order of score. For that purpose, systems often try to group together semantically equivalent answers. However, they hold several other semantic relations, such as inclusion, which are not considered, and candidates are mostly seen independently, as competitors.

Our hypothesis is that not just equivalence, but other relations between candidate answers have impact on the performance of a redundancy-based QA system. In this paper, we describe experimental studies to back up this hypothesis. Our findings show that, with relatively simple techniques to recognize relations, systems' accuracy can be improved for answers of categories NUMBER, DATE and ENTITY.

1 Introduction

Question-Answering (QA) systems automatically answer questions in natural language. To this end, a common approach to factoid questions relies on the direct extraction of candidate answers from document collections or the Web, while the selection of the final answer(s) is considered a ranking problem that benefits from the redundancy of the information sources. A score is attributed to each candidate, usually as a function of its frequency of occurrence, and the final answer(s) is/are selected from a list of candidates sorted in decreasing order of score, in which the top (or top N) is/are returned back to the user.

Approaches based on redundancy proved to achieve good results in QA, while requiring little linguistic processing, and have been extensively used since the early 2000's with the first works from Brill et al. [2001], Clarke et al. [2001] and Kwok et al. [2001]. However, and regarding answer selection, such strategies usually consider candidate answers as autonomous and independent entities, which is often not the case. For example, the candidate answers *Imola* and *Italy*

for the question “Where did Ayrton Senna have the accident that caused his death?” are related (since *Imola* is located in *Italy*) and probably they should not be dealt with separately.

In this context, Dalmas and Webber [2007] draw our attention to a new vision that sees candidate answers as allies, rather than competitors. Indeed, setting and using the semantic relations between candidate answers can potentiate the QA systems performance, by addressing a three-fold purpose:

1. allowing answer comparison and integration from various and heterogeneous sources;
2. improving answer selection; and,
3. allowing to build *better* answers.

In this paper, we address the problem of improving answer selection. Our hypothesis is that answer selection will benefit from the semantic relations detected between the candidate answers to a given question. For that, we part from the answer selection approach based on frequency and extend it with information about the semantic relations existing between answers. We show their impact on this phase of QA and, thus, in the overall performance of a redundancy-based QA system.

This paper is organized as follows: Section 2 describes related work; Section 3 presents the approach to answer selection based on semantic relations; Section 4 introduces the experimental setup and Section 5 details the experiments and results. The paper finishes in Section 6, with conclusions and future work directions.

2 Related Work

Although not too extensive, the literature in QA shows some examples of works that consider the semantic relations between candidate answers.

Relating candidate answers can be applied with the goal of comparing and integrating information chunks originated from different and heterogeneous sources. Normalization is often a step that proceeds the detection of relations. For instance, Moriceau [2005; 2006] compares and integrates answers of categories DATE and NUMBER and Chu-Carroll et al. [2003] uses named entity normalization on the candidates retrieved from two answering agents.

Regarding answer selection, it can be used when not many instances of the correct answer are found in the information

sources. Dalmas and Webber [2007] dealt with this issue and detected equivalences and inclusions between candidate answers to *where* questions. They organize answers in graphs and compute several features, like the semantic field, specificity and redundancy, that helps ranking the candidates. In our paper, we show the importance of such relations to a broader spectrum of questions. Moreover, and contrary to the work of Dalmas and Webber in which the candidate answers were synthetically generated, ours are the output of real-world QA systems. We consider the relations as a way to improve the results already achieved by redundancy, besides showing results when using the relations alone. Buchholz and Daelemans [2001] identified several situations where the success of the system can pass by the recognition of the relations between candidate answers: for instance, knowing that *cars* is a more important than *the pepper grinder which the company patented last century* allows a system to output a more appropriate answer to the question *What does the Peugeot company manufacture?*. The authors also referred to simple strategies that benefit from those relations, like, for example, selecting as final answer the one that includes most of the others.

Buchholz and Daelemans [2001] also considered relations in the context of the presentation of *complex* answers (answers composed by two or more simple answers). They identified several situations where the answers can be combined through the relations they hold, in order to build the system's answer. For instance, knowing that *Sirius* and *Dog Star* are different designations for the same entity, the system should choose and show the user a supporting sentence that contains both, to answer *"What is the brightest stater?"*.

Webber, Gardent and Bos [2003] originally introduced a set of four relations between answers, which were later presented by Moriceau [2005].¹ Both previous works consider the relations between correct answers, assuming a preceding filtering phase where the incorrect ones are discarded. In our work, on the contrary, we deal with relations between (correct and incorrect) candidate answers and discuss how they influence the selection of the correct answer. Relations between answers can be of:

Equivalence – if answers are consistent and entail mutually. Equivalence exists between lexicographically different answers that represent the same entity, namely:

- answers with notational variations. For instance, *Oct. 14, 1947* and *14th October, 1947* are equivalent answers for *"When did the test pilot Chuck Yeager break the sound barrier?"*;
- answers that rephrase others, like synonyms or paraphrases. For example, the question *"How did Jack Unterweger die?"* can be answered with *committed suicide* or *killed himself*.

Inclusion – if answers are consistent and differ in specificity, one entailing the other. Inclusion occurs between two candidates that represent different entities, in which one includes or subsumes the other, through:

- hypernymy, that is, the answers are in a *is-a* relation. For example, *"What animal is Kermit?"* can be answered with *frog* or *amphibian*;
- meronymy, that is, the answers are in a *part-of* relation. For example, *"Where did Ayrton Senna have the accident that caused his death?"*, in which *Imola, Italy*, and *Europe* are possible answers;
- membership, that is, the answers are in a *instance-of* relation. For example, *Edvard Munch* is a member of a *Norwegian Symbolist painter*, both possible answers to *"Who painted the 'Scream'?"*.

Aggregation – if answers are consistent, but not mutually entailing. In aggregation, all candidate answers are potentially correct and can be integrated in the form of a conjunction. For example, the question *"What is Kermit?"* can be answered with *frog* or *muppet*, or a conjunction of both: *frog and muppet*;

Alternative – if answers are not consistent or not mutually entailing. In the case of questions with unique answers, only one can be correct. For example, the question *"How is the Pope?"* can be answered with *ill* or *healthy*, but not with both. In the case of questions with multiple answers, all the alternatives may be distinct answers. For example, *twenty-eight* and *twenty-nine* are alternative answers to *"How many days are in February?"*.

3 Answer selection based on semantic relations

The basic answer selection approach relies on the frequency of occurrence of each candidate answer: the most frequent candidate answer is chosen as final answer. We extend this approach with the information about the semantic relations existing between candidate answers. Our strategy is composed by three steps and goes as follows:

1. The candidate answers are normalized to a canonical format, which allows a better comparison between them: DATE answers are set to the form *D01 M01 Y1900*, the SPEED and DISTANCE units are standardized (*miles an hour* and *mph* are set to *Unit_mph*) and numbers (written either in alphabetic and/or numeric characters) are reduced to its numeric version with one decimal place (*1 hundred* is converted to *100.0*);
2. Every two answers are compared (in a case insensitive comparison). If they are equal we assume them to be the same entity and the score of each answer is increased by one. Here, the score of each answer is its frequency of occurrence;
3. The relations that exist between every pair of candidate answers are considered. Again, we compare every pair of candidate answers and detect if a relation exists. The answers' scores are updated, depending on their relations with other candidates.

In this work, we used the typology of relations presented by Moriceau [2005]. We focused particularly on the relations of equivalence and inclusion. Here, we describe the techniques we used to detecting those relations between two candidate

¹In her Ph.D. thesis, Moriceau [2007] also mentions the relation of complementarity, defining it as existing between candidate answers that require inference mechanisms to solve a given question.

answers and how they interfere with the redundancy-based approach to select the final answer amongst a group of candidate answers.

3.1 Detecting Equivalence and Inclusion

One of the main concerns in redundancy-based QA is to recognize equivalence between answers despite of their surface form, for instance, that *July the 3th, 1983* is equal to *07-03-1983*. This notion of equivalence goes from notational variations or reference resolution to the more complex concept of paraphrase, and it is an open issue in QA research. The score attributed to a certain answer is tightly coupled with its frequency and of its equivalents, and ignores other types of relations. However, it should be clear that, for instance in the question “*Where did Ayrton Senna have the accident that caused his death?*”, the existence of the candidate answer *Imola* should boost the score of the candidate *Italy*, since they are connected through a relation of meronymy.

We use different techniques to recognize if two answers are related by equivalence or inclusion, depending on the semantics of the answer. These techniques vary from simple heuristics encoded in regular expressions, to linguistically informed techniques.

We manually created five regular expressions to detect equivalent candidate answers that refer to persons names. For instance, if a candidate A_1 is a string of alphabetic characters, we test if candidate A_2 matches the regular expression: “(Mrs?\.\?) | (Dr\.) | (Mister) | (Madame) A_1 ”. To cope with misspelled answers, we calculate the Levenshtein distance between every two answers, if it is lower than a certain threshold, we assume the answers to be equivalent.

In numeric answers, to deal with variation in numeric values, we assume that if the two answers differ by less than a threshold, then they are equivalent. Inclusion is detected with recourse to rules that test if a number is contained in an interval: for instance, 78.5 is included in *over 78.0*.

When it comes to dates, we recognize inclusion by testing if one answer string contains the other. For instance, *1 January 2011* includes *1 January, January 2011* and *2011*.

For other types of answers, and to reduce spelling variations, we apply lemmatization and assume answers to be equivalent if their lemmas are equal. To detect inclusion, we benefit from the knowledge present in WordNet [Fellbaum, 1998]. For each pair of lemmatized candidate answers, we query the WordNet for their most common sense. Afterwards, we search the tree of hypernyms of each candidate’s sense and check whether the other candidate is its ancestor. If so, a relation of inclusion exists.

After detecting the relations, one can build a directed graph of candidate answers, in which the answers are the nodes and the relations between them the edges. This approach was explored by Dalmás and Webber [2007], however their graph also contained information to represent entities extracted from the question. Our representation of candidate answers is independent from the input question, since we only take into account the answers. Figure 1 depicts the graph of candidate answers, achieved after the detection of equivalence and inclusion relations. Straight lines represent rela-

tions of equivalence and arrows represent relations of inclusion, in which the answer at the start of the arrow includes the answer at the end of the arrow.²

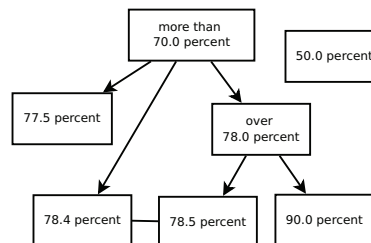


Figure 1: Relations between numeric candidate answers.

3.2 Updating the Answer’s Score

In the answer selection approach based uniquely on frequency candidates are scored independently and, for that reason, considered as autonomous. Here, the more popular an answer is, the more chances it has to be chosen as the final answer.

In the extended answer selection approach, candidate answers are scored according to the relations they hold with others, besides their frequency. The underlying assumption is that the correctness of an answer is influenced by the presence of the other answers with which it is related. Here, and recalling the graph representation of the candidate answers, the more connections an answer has, the more chances to be chosen as the final answer.

Being so, we implemented a scoring system, which updates the score of an answer A according to its frequency, to the number of answers equivalent to A , to the number of answers which include A and to the number of answers that are included by A . The strategy boils down to navigating the graph of candidate answers and adding to the score of each node a weight that depends on the number and type of edges that leave from and arrive to it. Finally, like in the frequency based approach, a list is built from the graph, ordered by decreasing value of score, and the topmost is chosen as final.

4 Experimental Setup

This section details the experimental setup, including the corpus of answers we used and the evaluation metrics.

4.1 Corpus of Answers

To our experiments we needed a large set of candidate answers to factoid questions, as well as their respective judgments as correct or incorrect. To deal with the difficulties in gathering such corpus from one unique system, we decided to use the freely available data from the Text REtrieval Conference (TREC) QA tracks. It contains not only the test questions, but also the answers given by the competing systems and judged as correct or incorrect by the human assessors of TREC. Being so, all our answers are originated in actual QA systems that used the same information sources to

²For simplicity reasons, we omitted transitive inclusions.

Category	# Questions		# Answers		
	All	Unsolv.	Correct	Inc.	All
HUMAN	423	16	2,723	9,304	12,027
ENTITY	224	6	970	5,154	6,124
NUMERIC	368	37	1,735	8,160	9,895
DATE	307	12	1,975	5,840	7,815
Total	1322	71	7,403	28,458	35,861

Table 1: Details of the corpus used in our experiments.

solve every question and their evaluation is trustworthy. We consider this to be a reliable and consistent corpus, that allows to mimic the behaviour of a real system, in line with a recent trend in QA – multi-stream QA – in which the output of several and different QA systems (streams) are combined to improve the answer accuracy [Télliez-Valero *et al.*, 2010].

We collected a total of 48,299 answers (10,635 judged as correct³ and 37,664 judged as incorrect) for 1,768 factoid questions, from the TREC QA tracks of the years of 2003 to 2007.⁴ Afterwards, we classified the questions according to Li and Roth’s two-layer question type taxonomy [Li and Roth, 2002]. We used an SVM classifier fed with features produced by a rule-based classifier which, in recent experiments, attained an accuracy of more than 95% and 90% for coarse- and fine-grained categories, respectively [Silva *et al.*, 2011]. Since the mappings between questions and answers are known, classifying questions is a way of categorizing the answers, allowing to filter out the answers of categories which are not the focus of the present work, as well as to use the proper techniques for recognizing relations according to the participating answers.

Our experiments focus on answers to factoid questions of categories ENTITY, NUMERIC and HUMAN. Thus, we discarded all answers of coarse-grained categories ABBREVIATION, LOCATION and DESCRIPTION and of the fine-grained category HUMAN:DESCRIPTION. We dealt with a total of 35,861 answers, in which 7,403 (20,64%) were judged as correct and 28,458 (79,36%) as incorrect. These answers belong to 1,322 different questions. Roughly, there is an average of 27 candidates per question, and 1 in every 5 answers is correct. Details of the corpus according to the category are presented in Table 1. Notice that the unsolvable questions are those for which none of its answers was judged as correct (questions that the system will not be able to solve).⁵

4.2 Evaluation Metrics

The hypothesis that motivates this work is that the relations between answers can improve the performance of QA systems, namely when selecting the final answer from a list of candidate answers.

Candidate answers are sorted in decreasing order of score

³Inexact and unsupported answers are considered correct.

⁴From 2003, questions have indication if they are factoids.

⁵Although in Li and Roth’s taxonomy the fine-grained category DATE belongs to the coarse-grained category NUMERIC, we separate them to differentiate the results achieved for dates or numbers. Being so, where we have NUMERIC, it should be understood as all questions/answers classified as NUMERIC except NUMERIC:DATE

(equally scored answers are positioned at the same rank) and we assume that any decision about picking answers from the group of candidate answers in a certain rank is uninformed, *i.e.*, it is random (uniformly distributed). Thus, for a question $q \in Q$, we measure the **probability of choosing the correct answer** at a rank r :

$$P(\text{Correct}A_{r,q}) = \frac{\#\text{Correct}CA_{r,q}}{\#CA_{r,q}} \quad (1)$$

In particular, if a question has M candidate answers at rank 1, in which N are correct, the system has a probability of N/M of choosing the correct answer.

The **number of correct questions at rank r** is the number of questions that choose the correct answer at least once from the top r ranked candidates; the **number of possibly correct questions at rank r** is the number of questions for which the probability of choosing the correct answer is more than 0.0 at least once in the top r ranked candidates.

The **accuracy** is measured as the total questions for which the probability of choosing the correct answer is 1.0 at rank 1, divided by the total number of questions in the test set.

$$\text{Acc.} = \frac{\#\{P(\text{Correct}A_{1,q}) = 1.0\}}{\#Q}, q \in Q \quad (2)$$

5 Experiments

Here we present the experiments and the achieved results.

5.1 Baseline

The baseline consists in selecting the most frequent candidate answer, without any other processing. That corresponds to the second step of the strategy presented in Section 3. Detailed results are shown in Table 2.

Results show an accuracy of 39.78% for the 1322 questions. Results are pushed down mostly because of the performance in the categories ENTITY and NUMERIC. When we consider the top 3 ranked answers, results improve in every category.

5.2 Normalizing Answers

Afterwards, we evaluated the impact of answer normalization in the baseline’s results. This corresponds to applying sequentially the steps 1 and 2 of the strategy presented in Section 3. In Table 3 we show the results.

When applying normalization before ranking the answers based only on frequency, results increase nearly 3% when compared with the baseline. Most of the 36 questions that are

Category	Accuracy	#CorrectQuestions@	
		1	3
HUMAN	44.92%	190	219
ENTITY	39.29%	88	97
NUMERIC	28.53%	105	119
DATE	46.58%	143	169
Overall	39.78%	526	604

Table 2: Results achieved when using only the answers frequency.

Category	Accuracy	#CorrectQuestions@	
		1	3
HUMAN	44.92%	190	219
ENTITY	40.18%	90	99
NUMERIC	32.07%	118	136
DATE	53.42%	164	187
Overall	42.51%	562	641

Table 3: Results achieved when using frequency, after normalizing the answers.

now correct at rank 1, belong to categories NUMERIC (13) and DATE (21). Regarding the 2 ENTITY questions that are now correct at rank 1, and since we did not use the normalizer in this category, they are two examples of misclassified questions, since both belong to the NUMERIC category. Considering the top ranked answers, results improve and are again better than those achieved by the baseline: 37 more questions have correct answer within the top 3.

5.3 Relating Answers

On a third experiment, we tested the impact of relating candidate answers. Regarding the number of relations between answers, with the techniques described in Section 3.1, we detected a total of 16,065 equivalences and 6,303 inclusions in 1,203 questions. We evaluated the performance of the answer selection approach based on semantic relations, which corresponds to all three sequential steps of the strategy presented in Section 3.

As it can be seen on Table 4, when relating candidate answers results increase nearly 10 and 7%, when compared to using only frequency, and frequency plus normalization, respectively. A total of 655 questions are correctly answered with the top ranked answer; considering the top 3 ranked answers, the difference is of 180 questions comparing with the baseline. Results suggest that the approach that takes the semantic relations into account better groups the correct answers in the top positions of the list of scored candidate answers. The largest difference in the number of correct answers was achieved in category NUMERIC, where a total of 4 rules for detecting equivalence and 8 for detecting inclusion were the reason for 44 and 69 more correct questions at rank 1 and 3, respectively. In the category ENTITY this difference was the smallest. We consider that this happened because ENTITY is too broad a category, that covers very diverse questions, and probably it requires other techniques for detection relations. Overall, results confirm the applicability

Category	Accuracy	#CorrectQuestions@	
		1	3
HUMAN	52.01%	220	265
ENTITY	42.41%	95	110
NUMERIC	44.02%	162	205
DATE	57.98%	178	204
Overall	49.55%	655	784

Table 4: Results achieved when using answer selection based on semantic relations.

Category	No Equivalence #CorrectQuestions@		No Inclusion #CorrectQuestions@	
	1	3	1	3
HUMAN	190	219	220	265
ENTITY	96	113	95	109
NUMERIC	140	151	158	203
DATE	178	204	164	187
Overall	590	687	637	764

Table 5: Results achieved when not using one of the relations.

of semantic relations for a broad variety of questions. As two examples of correct answers in the top 1 ranked, *sheep* was selected as final answer in a question whose candidates were related as follows: “*animal includes sheep*” and “*sheep Dolly equivalent to sheep*”. Also, the candidate *D13 M07* allowed the correct answer *D13 M07 Y1999* to be better scored than the wrong answer *D03 M11 Y1999*.

We measured the impact of normalization⁶. When this step is bypassed, the accuracy drops to 47.88%, with 633 correctly answered questions at rank 1. That is, 22 questions were correct due to normalization. We also assessed the influence of each relation on answer selection, while keeping normalization. Table 5 shows that results deteriorate regardless of the removed relation. Moreover, and although equivalence has a bigger impact, when inclusion is withdrawn, the total number of correct questions lower in about 20 questions. It is interesting to analyse the results achieved for categories ENTITY and NUMERIC.⁷ Regarding the latter, results show that both relations contribute differently for the best marks achieved. Concerning the former, and in contradiction with the other categories, equivalence seems to penalise the achieved results. Indeed, when no equivalence is used the number of correct answers is higher; however, when equivalence is used without inclusion, results are still better than the frequency-based approach with normalization. We consider these results an evidence of the possible influence of semantic relations on the overall results of a redundancy-based QA system.

Table 6 compares results achieved in the three previous experiments: frequency-based selection (baseline), normalization plus frequency and normalization plus frequency plus semantic relations. It presents the number of correct questions and of possibly correct questions at rank r ($1 \leq r \leq 5$).

In **i** and **ii**, for all questions at least one correct answer can be found in the top 5 ranked candidates (recall that the number of unsolved questions is 71 and the total number of questions 1,322). However, results suggest that the correct answers are disperse in the list and their score is not enough to distinguish them from the incorrect answers. Moreover, without any further information, choosing the correct answer at a rank r is a random decision whose performance degrades when increasing the rank. In **iii**, the ranked list of candidate answers is, in overall, longer (not all questions have a correct answer at the rank 5). Results show that a more fruitful selection of the final answer can be made at every rank, since

⁶These results are not present in the table.

⁷Categories HUMAN and DATE are only affected by equivalence or inclusion, respectively.

	Rank				
	1	2	3	4	5
i Frequency (Baseline)					
#CorrectQuestions	526	590	604	605	605
#PosCorrectQuest	785	1093	1231	1250	1251
ii Normalization + Frequency					
#CorrectQuestions	562	628	641	643	643
#PosCorrectQuest	787	1090	1224	1250	1251
iii Normalization + Frequency + Relations					
#CorrectQuestions	655	742	784	810	822
#PosCorrectQuest	778	947	1077	1173	1208

Table 6: Comparison between the previous experiments.

the number of correct questions is always superior to those in **i** and **ii**, regardless of the rank.

5.4 Further Experiments and Results

We simulated the behaviour of a system when the answers are not redundant and only one instance of each candidate exists. In this case, the answers are not dependent on their frequency, just on the relations with other answers. If all candidate answers have the same frequency (1), and without other information, a QA system would probably choose randomly amongst all answers. With the approach to answer selection based on semantic relations, and disregarding the frequency, an accuracy of 33.89% was achieved, with 448 questions correctly answers at the top 1 rank. That is, a difference of less than -6% compared with the baseline. Results without normalization drop to 431 correct questions at the first rank. This lead us to conclude that, although answer redundancy is a good measure of the correctness of an answer, when this property is not available, using normalization with the semantic relations between answers seems to be a good substitute.

6 Conclusions and Future Work

We presented an approach to answer selection in QA that takes into account not only the candidate answers' frequency, but also the relations they hold with other candidate answers. Using a limited set of heuristics, encoded mostly in the form of regular expressions, as well as linguistic knowledge from WordNet, we build a graph which we traverse to update the score of every answer. With this approach, that uses mostly information recovered from the answer, we could boost the accuracy of the baseline in nearly 10%. We presented a detailed evaluation and we discussed the impact of frequency, normalization and the semantics relations for the purpose of ranking candidates and selecting the final answer.

As future work, we intend to improve the current techniques and explore others for detecting relations between candidate answers, like, for instance, paraphrase recognition. It is also in our plans to learn the optimal weights to update an answer's score depending on the relations it holds with others.

Acknowledgments

This work was supported by FCT (INESC-ID multianual funding) through the PIDDAC Program funds, and

also through the project CMU-PT/0005/2007. Ana Cristina Mendes is supported by a PhD fellowship from Fundação para a Ciência e a Tecnologia (SFRH/BD/43487/2008).

References

- [Brill *et al.*, 2001] E. Brill, J. Lin, M. Banko, S. Dumais, and A. Ng. Data-intensive question answering. In *Proc. 10th Text REtrieval Conference (TREC)*, pages 393–400, 2001.
- [Buchholz and Daelemans, 2001] S. Buchholz and W. Daelemans. Complex answers: a case study using a www question answering system. *Nat. Lang. Eng.*, 7:301–323, 2001.
- [Chu-Carroll *et al.*, 2003] J. Chu-Carroll, K. Czuba, J. Prager, and A. Ittycheriah. In question answering, two heads are better than one. In *Proc. NAACL'03*, pages 24–31. ACL, 2003.
- [Clarke *et al.*, 2001] C. Clarke, G. Cormack, and T. Lynam. Exploiting redundancy in question answering. In *Proc. of SIGIR*, pages 358–365. ACM Press, 2001.
- [Dalmas and Webber, 2007] T. Dalmas and B. Webber. Answer comparison in automated question answering. *Journal of Applied Logic*, pages 104–120, 2007.
- [Fellbaum, 1998] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
- [Kwok *et al.*, 2001] C. Kwok, O. Etzioni, and D. Weld. Scaling question answering to the web. *ACM Trans. Inf. Syst.*, 19:242–262, 2001.
- [Li and Roth, 2002] X. Li and D. Roth. Learning question classifiers. In *Proc. 19th Int. Conf. Computational linguistics*, pages 1–7. ACL, 2002.
- [Moriceau, 2005] V. Moriceau. Answer generation with temporal data integration. In *Proc. 10th European Workshop Nat. Lang. Generation (ENLG-05)*, pages 197–202, 2005.
- [Moriceau, 2006] V. Moriceau. Numerical data integration for cooperative question-answering. In *Proc. Workshop KRAQ'06 on Knowledge and Reasoning for Language Processing*, KRAQ '06, pages 42–49. ACL, 2006.
- [Moriceau, 2007] V. Moriceau. *Intégration de données dans un système question-réponse sur le Web*. PhD thesis, Université Paul Sabatier, 2007.
- [Silva *et al.*, 2011] João Silva, Luísa Coheur, Ana Mendes, and Andreas Wichert. From symbolic to sub-symbolic information in question classification. *Artificial Intelligence Review*, 35:137–154, 2011. 10.1007/s10462-010-9188-4.
- [Téllez-Valero *et al.*, 2010] A. Téllez-Valero, M. Montes-Gómez, L. Villaseñor Pineda, and A. Peñas. Towards Multi-Stream Question Answering using Answer Validation. *Informatica. Special Issue on Computational Linguistics and its Applications*, 34(1), 2010.
- [Webber *et al.*, 2003] B. Webber, C. Gardent, and J. Bos. Position statement: Inference in question answering. In *Proc. LREC Workshop on Question Answering: Strategy and Resources*, pages 19–25, 2003.