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	1096	1224	1250	1251
iii Normalization + Frequency + Relations					
#CorrectQuestions	665	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 multiannual 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. 10^{th} 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.