Category	Accuracy	#CorrectQuestions@ 1	3
HUMAN	44.92%	190	219
ENTITY	40.18%	96	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.

	No Equivalence #CorrectQuestions@		No Inclusion #CorrectQuestions@	
Category	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 normalization[6]. 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

[6]These results are not present in the table.

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