The LIMSI participation to the QAst 2009 track

Guillaume Bernard, Sophie Rosset, Olivier Galibert, Eric Bilinski, Gilles Adda,
Spoken Language Processing Group, LIMSI-CNRS, B.P. 133, 91403 Orsay cedex, France
{firstname.lastname}@limsi.fr

Résumé

We present in this paper the three LIMSI question-answering systems on speech transcripts which participated to the QAst 2009 evaluation. These systems are based on a complete and multi-level analysis of both queries and documents. These systems use an automatically generated research descriptor. A score based on those descriptors is used to select documents and snippets. Three different methods are tried to extract and score candidate answers, and we present in particular a tree transformation based ranking method. We participated to all the tasks and submitted 30 runs (for 24 sub-tasks). The evaluation results for manual transcripts range from 27% to 36% for accuracy depending on the task and from 20% to 29% for automatic transcripts.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing ; H.3.3 Information Search and Retrieval ; H.3.4 Systems and Software ; H.3.7 Digital Libraries

General Terms

Measurement, Performance, Experimentation

Keywords

Question answering, speech transcriptions

1 Introduction

The Question Answering on Speech Transcripts track of the QA@CLEF task provides an opportunity to evaluate the specificity of speech transcriptions. In this paper, we present the work carried out on the QA system developed at LIMSI for the QAst evaluation. We especially describe an answer re-ranking method used in this system.

For the QAst 2009 evaluation [7], 3 main tasks are defined :

– T1, QA in English European Parliament Plenary sessions
– T2, QA in Spanish European Parliament Plenary sessions
– T3, QA in French Broadcast News

In the previous QAst evaluations [6], the questions were created by the evaluators from the documents. This year, the objective was to have more spontaneous questions. Native speakers were requested to read

fragments of doucments and ask, using speech, questions about information related to but not content in the texts fragments.

For each of the tasks, four versions of the data collection were provided, consisting of one manual transcriptions and three different automatic transcription. Two different sets of questions were provided, one consisting of written questions and the other of manually transcribed semi-spontaneous oral questions. In total a minimum of 8 runs were expected per task, for a total of 24. LIMSI participated to the three tasks. Three systems were tested. Their main architecture is identical and they differ only in the answer scoring method :

– Distance-based answer scoring (primary method)
– Answer scoring through bayesian modeling
– Tree transformation-based answer re-ranking

The first method is used on all three tasks, the second is used on the T1 and T2 tasks and the third on the T3 task.

The section 2 presents the common architecture and the answer scoring methods. The section 3 is split into three parts : the description of the training and development data (section 3.1, a quick evaluation of the difficulty of the task (section 3.2), and finally the results of the three systems on the development and test data (sectiob 3.3). We compare these results to those obtained in the QAst 2008 evaluation.

2 The LIMSI QA systems

The common architecture is identical to the systems used in the previous evaluations and is fully described in [4].

The same complete and multilevel analysis is carried out on both queries and documents. To do so, the query and the documents (which may come from different modalities – text, manual transcripts, automatic transcripts) are transformed into a common representation. This normalization process converts *raw* texts to a form where words and numbers are unambiguously delimited, punctuation is separated from words, and the text is split into sentence-like segments. Case and punctuation are reconstructed using a fully cased, punctuated four-gram language model [1] applied to a word graph covering all the possible variants (all possible punctuations permitted between words, all possible word cases). The general objective of this analysis is to find the bits of information that may be of use for search and extraction, called *pertinent information chunks*. These can be of different categories : named entities, linguistic entities (e.g., verbs, prepositions), or specific entities (e.g., scores). All words that do not fall into such chunks are automatically grouped into chunks via a longest-match strategy. The full analysis comprises some 100 steps and takes roughly 4 ms on a typical user or document sentence. The analysis identifies about 300 different types of entities. The analysis is hierarchichal, resulting in a set of trees. Both answers and important element of the questions are supposed to be annotated as one of these entities.

The first step of QA system itself is to build a search descriptor (SD) that contains the important elements of the question, and the possible answer types with associated weights. Some elements are marked as *critical*, which makes them mandatory in future steps, while others are *secondary*. The element extraction and weighting is based on an empirical classification of the element types in importance levels. Answer types are predicted through rules based on combinations of elements of the question.

Documents are selected using this SD. Each element of the document is scored with the geometric mean of the number of occurrences of all the SD elements that appear in it, and sorted by score, keeping the *n*-best.

Snippets are extracted from the document using fixed-size windows and scored using the geometrical mean of the number of occurrences of all the DDR elements that appear in the snippet, smoothed by the document score.

2.1 Distance-based answer scoring

In each snippet, all the elements whose type is one of the predicted possible answer types are candidate answers. A score $S(r)$ is associated to each candidate answer *r* :

$$S(r) = \frac{\sum_{a \in A_r} (w(a) \max_{E_a} \sum_{(e,l) \in E_a} \frac{w(l)}{(1+d(e,a))^{\alpha}})^{1-\gamma} S_p(a)^{\gamma}}{C_d(r)^{\beta} C_p(r)^{\delta}}$$

$w(l)$ = line weight $w(a)$ = answer weight
$d(e, a)$ = element-answer distance
E_a = set of SD elements for instance a
A_r = set of instances of the answer candidate r
$S_p(a)$ = score of the snippet including a
$C_d(r)$ = instance count of r in the documents
$C_p(r)$ = instance count of r in the snippets
$\alpha, \beta, \gamma, \delta$ = tuning variables

2.2 Answer scoring through bayesian modeling

We tried a preliminary method of answer scoring built upon a bayesian modeling of the process of estimating the quality of an answer candidate. This approach relies on multiple elementary models including element co-occurrence probabilities, question element appearance probility in the context of a correct answer and out of context answer probability. This is a very preliminary work.

2.3 Tree transformation-based answer re-ranking

Our second approach for the T3 task is built upon the results of the primary system. We stated that the method for finding and extracting the best answer to a given question in 2.1 is based on redundancy and distances between candidate answers and elements of the question. While this approach gives good results, it also has some limitations. Mainly, it does not take into account the structure of the snippet and the relations between the different critical elements detected.

Relations between the elements of the text fragments are needed to represent the information stated in the documents and the questions. However, most of the systems use complex syntactic representations which are not adapted to handle oral fragments[2]. However, some systems[5, 3] show that it is possible to identify local syntactic and semantic relations by using a segmentation of the documents into segments (chunks) and then detecting the relations between these segments.

From these conclusions, we defined a re-ranking method which computes a score for each of the answers to a question. That method takes as input the question tagged by the analysis module, the answers found by the

answer extraction module, and the best snippets associated to each answer. The analysis trees of the question and the snippets are segmented into chunks, and relations are added between these chunks.

For each evaluated answer, the method compares the structure of the question with the snippet of the answer. The system tries to match the structure of the question by moving the chunks of the snippets with similar elements. The relations are used in these moves and allow the system to compute the score of the answer.

This system uses two sub-modules, the segmenting and annotation module and the relation labelling module. The questions and the snippets are processed through these modules, and then the tree transformation system computes the similarity score of each answers.

2.3.1 Segmentation and annotation module

The definition of the segmentation formalism was led by its use for the relation labelling module. We think that verbs have an important role in the structure of a sentence. Therefore, we have defined two types of chunks : verbal chunks (VC) and general chunks (GC). The general chunks can be divided into several subtypes : temporal (TC), spatial (SC) and question markers (QMC). Below is an exemple of a segmented sentence, *"The Ebola virus was identified in 1976"*.

"[GC] The Ebola virus [/GC] [VC] was identified [/VC] [TC] in 1976 [/TC]."

The segmentation and annotation module uses a Conditional Random Fields (CRF) based approach. Two models were generated : one for the documents, and one for the questions. We used the following features : analysis module of the main architecture and a Part Of Speech annotation. Two training corpus were used, one for the documents and one for the questions.

2.3.2 Relation labelling module

The aim of the relations is to represent the dependances between the chunks of the questions and the chunks of the snippets.

The relations are oriented and non-exclusive, ie there can be multiple relations between the same two chunks. For the moment, five relations are defined, which are described below.

Noun modifier relation ; this relation represents the dependance between two chunks containing noun groups, as in the following sentence : *"[GC] Steven Spielberg [/GC] [VC] is [/VC] [GC] the man [/GC] [GC] with the glasses [/GC]"*. In this example, there is a **noun modifier relation** between *"the man"* and *"with the glasses"*.

Verb to member relation ; this relation represents the dependance between a verbal chunk and the chunks containing its members. The members of a verb are its subject and its objects. In the following sentence, *"[GC] The Ebola virus [/GC] [VC] was identified [/VC] [TC] in 1976 [/TC]."*, there are two **verb to member relations** between the verbal chunk *"was identified"* and the two chunks *"in 1976"* and *"The Ebola virus"*.

Member to verb relation ; this type of relation is the same as the previous one, except this relation goes from the member to the verb.

Temporal relation ; this relation represents the dependance between a temporal chunk and another chunk.

In the following sentence, *"[GC] The Ebola virus [/GC] [VC] was identified [/VC] [TC] in 1976 [/TC]."*, there are two **temporal relations** between the temporal chunk *"in 1976"* and the two chunks *"was identified"* and *"The Ebola virus"*.

Spatial relation; this type of relation is the same as the temporal relation, except that it concerns spatial chunks.

To label the different relations of each of the chunk of the documents and the questions, we use a rule-based system. Each type of relation has an associated rule, with the following parameters : the types of the chunks on which the rule applies, the types of chunks who can be in relation, the direction of the rule, and the context of application of the rule. Here is an exemple of the rule for *temporal relations* :

temporal relation : {TC} {GC | VC | SC | QMC} {LEFT | RIGHT} {TC}

This rule means that we add a *temporal relation* between a chunk of TC type and a chunk of GC, VC, SC or QMC type. The target chunk can either be at the left or a the right of the temporal chunk. The relation is not allowed to cross over another temporal chunk.

2.3.3 Text transformation module

As we said previously, before trying to transform the snippet into the question, the system finds the similarities between the chunks of the snippet and the chunks of the question. To find the similarities, we use the following information : lemma form, synonyms and morphological derivations. The system defines anchor points between comparable chunks.

With these points, the system transforms the snippet into the question by using three types of operations : inserting a chunk, deleting a chunk and substituting a chunk. These types of operation are applied in a certain order.

First, the system generates one substituting operation for each anchor point, and compute its cost. It depends on two values : the substitution cost and the displacement cost. The subtitution cost is computing by making the sum of a per word cost for each important word which is not found in the question. We decide wether a word is deemed important based on its type given by the analysis module. For example, verbs and nouns are important but determinants are not. The per word cost has been set empirically. The displacement is seen as a sequence of permutations between adjacent chunks. Each permutation has a cost depending on the relation between the two chunks and their types. Then, the system finds the sequence of substitutin operations with the lowest total score, which results in a similar structure between the question and the snippet. To finish the transformation, the remaining chunks are deleted, and the missing ones from the question are inserted. The sequence of operations with the smallest total cost measures the similarity between the question and the snippet, and by comparing these similarity scores a new ranking is computed.

The figure 1 shows an example of how the transformation works. The relations are not shown for clarity. We evaluate the answer *"Northern Ireland"* for the question *"What country is Annetta Flanigan from ?"*. The snippet of the answer is *"One captive is Annetta Flanigan from my constituency of Northern Ireland"*.

As you can see on the figure, the system find three anchor points between the chunks of the question and those of the text fragment. The colours show the anchors between the chunks. Using these anchors, the system generates the operations. For this example, a list of transformation could be :

– Moving chunk *"Annetta Flanigan"* next to chunk *"of Northern Ireland"*

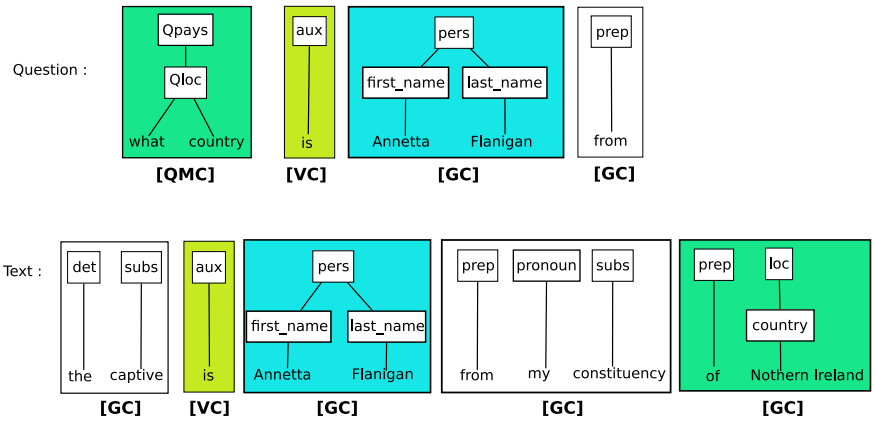

FIGURE 1 – Transformation exemple between the question ”What country is Annetta Flanigan from ?” and the snippet ”One captive is Annetta Flanigan from my constituency of Northern Ireland” containing the answer ”Northern Ireland”

– Deleting *”the captive”* and *”from my constituency”*

We think that verbs in a sentence are important to the signification and the structure of this sentence. Thus, we do not allow a permutation between a verbal chunk and another general chunk. That is why in this example we do not allow a permutation between *”Northern Ireland”* and *”is”*. Also, we do not insert the preposition *”from”* because it is not a chunk of the question associated to a critical element of the search descriptor (SD). This sequence of operations allows the system to compute a new score for this answer.

3 Evaluation

3.1 Training and Development data

Each main task had a two 50 questions development corpus, one of manual transcriptions of spontaneous spoken questions and one of written reformulations of these questions.

An overview of the different corpus used can be viewed in table 1. The numbers between the parenthesis indicate the number of documents of the corpus.

As part of the training data, we used the corpus of reformulated questions we developped last year in addition to the official development corpus and the test data from QAst 2008.

Task	Off. Dev.	Supp. Dev.
T1	2×50 (6)	377
T2	2×50 (6)	317
T3	2×50 (18)	450

TABLE 1 – The corpus. Off. Dev : the official development data ; Supp. Dev : the QAst08 development and test data and reformulated questions based on the QAst08 development

Since the second approach for the T3 task, the re-ranking method, does not yet give better results on the

whole development corpus, we decided to use it only on certain type of questions. In Table 2, the LIMSI1 system is the distance-based method, and the LIMSI2-T3 is the re-ranking method. We found that this method gets better results on questions with a lot of search elements present in the search descriptor. The corpus of questions evaluated in this table is a fusion between the official and the supplementary development corpus. We can see that on questions with at least 5 search elements, the LIMSI2 on the T3 task gets better results than the LIMSI1 system. Thus, we decided that the re-ranking approach will only be applied on questions with at least 5 search elements.

	LIMSI2-T3			LIMSI1			
#E.	MRR	Acc	#Correct	MRR	Acc	#Correct	#Questions
1	0.62	48.6	53	0.71	67.0	73	109
2	0.56	42.2	73	0.66	61.8	107	173
3	0.74	67.4	145	0.79	77.9	166	215
4	0.72	65.5	74	0.79	77.9	88	113
5	**0.73**	**65.5**	38	0.71	60.3	35	58
6	**0.85**	**85.7**	18	0.81	76.0	16	21
7	0.60	60.0	3	0.60	60.0	3	5

TABLE 2 – Results on the development data classified by number of search elements in the search descriptor, with #E being the number of search elements. LIMSI2-T3 is the re-ranking method and LIMSI1 the distance-based method.

3.2 Task difficulty evaluation

As stated in Section 1, the procedure for building the question corpus has changed this year. We try to evaluate whether the difficulty of the task had changed as a result.

Mainly, we wanted to compare the differences between the development corpus of QAst08 and QAst09. Therefore, we evaluated for each question of the two corpus the distance between the elements of the question and the answer in the documents containing the correct answer. For each questions, we computed four distance scores : the number of words, the number of nodes of the analysis module, the number of chunks and the number of sentences. Each score is an average of the distance of each element of the question from the answer. This evaluation was made on the corpus of the T3 main task (French corpus). The Table 3 shows the results of this evaluation.

	Words		Nodes		Chunks	
	Mean	SD	Mean	SD	Mean	SD
QAst09 written development corpus	27	52	47	17	10	20
QAst09 spoken development corpus	28	52	47	22	10	20
QAst08 development corpus	14	20	13	23	5	7
QAst08 reformulated corpus	18	26	15	25	5	9

TABLE 3 – Mean and standard deviation (SD) of the distance between correct answers and elements of the questions in various units

This table shows some differences between the development corpus of QAst09 and QAst08. We see that the mean distance is roughly doubled in the QAst09 development corpus compared to the previous year. While such a difference is significant in absolute terms, we do not think that it by itself fully justifies a large difference in task difficulty. We need to also analyse the impact of lexical variations between the questions and the documents contents.

3.3 Results

3.3.1 General results on manual transcripts

The results for the three tasks on manual transcribed data are presented in tables 4 to 6, with all the question types evaluated. For each task, two systems were used. There is also a difference between the LIMSI2 system in the T1 and T2 tasks (English and Spanish) and LIMSI2 system in the T3 task (French). For each case, only the Factual Answer Extraction procedure is changing : in LIMSI1, it uses a scoring of all the candidates of appropriate types given the question classification. In LIMSI2 for the T1 and T2 tasks, it uses the bayesian method explained before, and in the T3 task the tree transformation re-ranking method. As stated before, the LIMSI2 system in the T3 task is not used on all the questions, but only the questions with a lot of search elements.

System	Questions	Test 09		
		MRR	Acc	Recall
LIMSI1	Written	0.36	27%	53%
	Spoken	0.33	23%	45%
LIMSI2	Written	0.32	23%	45%
	Spoken	0.27	19%	41%

TABLE 4 – Results for task T1, English EPPS, manual transcripts (75 factual questions and 25 definitional ones).

System	Questions	Test 09		
		MRR	Acc	Recall
LIMSI1	Written	0.45	36.0%	61%
	Spoken	0.45	36.0%	62%
LIMSI2	Written	0.34	24.0%	49%
	Spoken	0.34	24.0%	49%

TABLE 5 – Results for task T2, Spanish EPPS, manual transcripts (44 factual questions and 56 definitional ones).

System	Questions	Test 09		
		MRR	Acc	Recall
LIMSI1	Written	0.39	28.0%	60%
	Spoken	0.39	28.0%	59%
LIMSI2	Written	0.38	27.0%	60%
	Spoken	0.39	28.0%	59%

TABLE 6 – Results for the T3 task, French Broadcast News, manual transcripts (68 factual questions and 32 definitional ones).

3.3.2 General results on automatic transcripts

The results obtained on the three tasks in automatically transcribed data are presented in tables 7 to 9. With the automatic transcripts, only the LIMSI1 system is used.

ASR	System	Questions	Test 09		
			MRR	Acc	Recall
ASR_A	LIMSI1	Written	0.31	26.0%	42%
10.6%		Spoken	0.30	25.0%	41%
ASR_B	LIMSI1	Written	0.25	21.0%	32%
14.0%		Spoken	0.25	21.0%	33%
ASR_C	LIMSI1	Written	0.24	21.0%	31%
24.1%		Spoken	0.24	20.0%	33%

TABLE 7 – Results for task T1, English EPPS, automatic transcripts (75 factual questions and 25 definitional ones).

ASR	System	Questions	Test 09		
			MRR	Acc	Recall
ASR_A	LIMSI1	Written	0.32	27.0%	42%
11.5%		Spoken	0.31	26.0%	41%
ASR_B	LIMSI1	Written	0.29	25.0%	37%
12.7%		Spoken	0.29	25.0%	37%
ASR_C	LIMSI1	Written	0.28	23.0%	37%
13.7%		Spoken	0.28	24.0%	37%

TABLE 8 – Results for task T2, Spanish EPPS, automatic transcripts (44 factual questions and 56 definitional ones).

ASR	System	Questions	Test 09		
			MRR	Acc	Recall
ASR_A	LIMSI1	Written	0.37	29.0%	52%
11.0%		Spoken	0.37	29.0%	50%
ASR_B	LIMSI1	Written	0.32	27.0%	40%
23.9%		Spoken	0.30	25.0%	38%
ASR_C	LIMSI1	Written	0.28	23.0%	38%
35.4%		Spoken	0.27	22.0%	35%

TABLE 9 – Results for the T3 task, French Broadcast News, manual transcripts (68 factual questions and 32 definitional ones).

3.3.3 Analysis of the results

Tables 4 to 6 show a great loss between the recall and the accuracy of our systems. The LIMSI1 system gives a bad answer on half of the questions with the good answer in the candidates answers, and it is worst for the LIMSI2 system on the T1 and T2 tasks. The LIMSI2 system on the T3 task gives almost the same results that the LIMSI1 system by applying it only on a small set of questions, as stated previously. A study of the results of this system is showed next. Nevertheless, we can see that there are almost no differences between written and spoken questions. LIMSI2 system on the T1 and T2 tasks is a preliminary version that gives interesting results. As such, we are going to improve it. LIMSI2 system on the T3 task still needs work to improve it.

For the results obtained on the three different automatic speech transcription, as showed in tables 7 to 9, we

can see that they are lower than the results of the manual transcriptions.

We show in table 10 the results obtained by the LIMSI on each task. We also show the best results of all the participants systems in column *Best* for each task. Except on the T1 Manual and the T1 ASR_A, the LIMSI obtains the best results. It should be noted that we were the only participants in the T3 task.

		T1		T2		T3	
Sub-Task	**Question**	**Acc.**	**Best**	**Acc.**	**Best**	**Acc.**	**Best**
Manual	Written	27.0%	28.0%	36.0%	-	28.0%	-
	Spoken	23.0%	26.0%	36.0%	-	28.0%	-
ASR_A	Written	26.0%	-	27.0%	-	29.0%	-
	Spoken	25.0%	-	26.0%	-	29.0%	-
ASR_B	Written	21.0%	-	25.0%	-	27.0%	-
	Spoken	21.0%	-	25.0%	-	25.0%	-
ASR_C	Written	21.0%	25.0%	23.0%	-	23.0%	-
	Spoken	20.0%	25.0%	24.0%	-	22.0%	-

TABLE 10 – Results obtained by the LIMSI on the QAst 2009 evaluation.

Table 11 shows the results obtained by each system for the manual sub-tasks on the T1 and T2 tasks. The evaluated corpus are the development and the test corpus of QAst 2009 on both written and spoken questions, and the development and test corpus of QAst 2008, on written questions. As stated before, the LIMSI1 system used the distance-based approach, and the LIMSI2-T1 and LIMSI2-T2 the bayesian approach.

	T1				T2			
	LIMSI1		LIMSI2-T1		LIMSI1		LIMSI2-T2	
Corpus	MRR	Acc	MRR	Acc	MRR	Acc	MRR	Acc
W. Test09	0.36	27%	0.32	23%	0.45	36%	0.34	24%
S. Test09	0.33	23%	0.27	19%	0.45	36%	0.34	24%
W. Dev09	0.37	32%	0.21	10%	0.54	48%	0.37	26%
S. Dev09	0.39	34%	0.22	10%	0.52	45%	0.42	32%
W. Dev08	0.80	78%	0.59	50%	0.68	58%	0.57	42%
W. Test08	0.55	52%	0.38	32%	0.62	56%	0.52	44%

TABLE 11 – Results obtained on each system for the manual tasks.

Table 12 compared the results on the T3 task between the LIMSI1 system and the LIMSI2-T3 system. We show two sets of results for the LIMSI2 system : those obtained where all the questions of the corpus are re-ranked (LIMSI2-T3), and those obtained where only the questions with 5 or more search elements are re-ranked (LIMSI2-T3-SE).

As we can see, there is an huge loss between the QAst08 corpus and the test and development corpus of QAst09. One reason for these results could be the new methodology used to build the questions corpus. As stated in section 3.1, the distances between the elements of the question and the answer are greater in the development corpus of QAst09. The greater distance between an answer and its associated question elements does not seem to be the only cause of these results. In particular, we expect lexical variations between the questions and the elements as found in the documents to also have play a significant role.

Table 12 shows that the re-ranking of the questions with 5 or more search elements allows the LIMSI2-T3 system to get almost the same results than the LIMSI1 system. We can see that it also gets better results on the development corpus from QAst 2008. While these results are interesting, as stated before this approach needs to be improved.

	T3					
	LIMSI1		LIMSI2-T3		LIMSI2-T3-SE	
Corpus	MRR	Acc	MRR	Acc	MRR	Acc
W. Test09	0.39	28%	0.24	18%	**0.38**	**27%**
S. Test09	0.39	28%	0.24	17%	**0.39**	**28%**
W. Dev09	0.44	40%	0.25	16%	0.44	40%
S. Dev09	0.44	36%	0.26	18%	0.42	34%
W. Dev08	0.81	76%	0.68	58%	**0.85**	**80%**
W. Test08	0.57	50%	0.50	40%	0.57	50%

TABLE 12 – Results obtained on each system for the manual tasks.

We evaluated the questions with 5 or more search elements which were re-ranked by the LIMSI2-T3 system. Of the ten questions of the written question corpus with that many search elements, six did not have the correct answer ithin the candidates answers and one was a NIL question. Of the remaining three, one was answered correctly by both systems, one was answered correctly by the LIMSI1 but not the LIMSI2-T3 one. And the correct answer for the last question was not found by either of the systems.

4 Conclusion

In this paper, we presented the LIMSI question-answering systems on speech transcripts which participated to the QAst 2009 evaluation. These systems obtained state-of-the-art results on the different tasks and languages and the accuracy ranged from 27% for English to 36% for Spanish data). The results of the T1 ans T3 systems show a significant loss of results compared to the 2008 evaluation (6% for T1 and 17% for T3 in accuracy) inspite of the improvements og the systems. It can be explained by the new methodolgy used to build the questions corpus. A deeper analysis is ongoing to understand the origins of this loss.

Acknowledgments

This work has been partially financed by OSEO under the Quaero program.

Références

[1] D. Déchelotte, H. Schwenk, G. Adda, and J.-L. Gauvain. Improved machine translation of speech-to-text outputs. Antwerp. Belgium, 2007.

[2] P. Paroubek, A. Vilnat, B. Grau, and C. Ayache. Easy, evaluation of parsers of french : what are the results ? In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 2480–2486, Marrakech, Morocco, 2008.

[3] S. Pradhan, W. Ward, K. Hacioglu, J. H. Martin, and D. Jurafski. Semantic role labeling using different syntactic views. pages 581–588, Ann Arbor, USA, 2005.

[4] S. Rosset, O. Galibert, G. Bernard, E. Bilinski, and G. Adda. The limsi participation to the qast track. In *Working Notes of CLEF 2008 Workshop*, Aarhus, Denmark, September 2008.

[5] T. Sakai, Y. Saito, Y. Ichimura, M. Koyama, T. Kokubu, and T. Manabe. Askmi : A japanese question answering system based on semantic role analysis. In *Proceedings of RIAO 2004*, Avignon, 2004.

[6] J. Turmo, P. Comas, L. Lamel, S. Rosset, N. Moreau, and D. Mostefa. Overview of qast 2008 - question answering on speech transcriptions. In *CLEF 2008 Workshop*, Aarhus, Denmark, 2008.

[7] J. Turmo, P. Comas, S. Rosset, O. Galibert, N. Moreau, D. Mostefa, P. Rosso, and D. Buscaldi. Overview of qast 2009 - question answering on speech transcriptions. In *CLEF 2009 Workshop*, Greece, Corfu, 2009, to appear.