Named Entity Recognition in Question Answering of Speech Data



Diego Mollá **Menno van Zaanen** **Steve Cassidy**

Centre for Language Technology, Macquarie University, Sydney, Australia

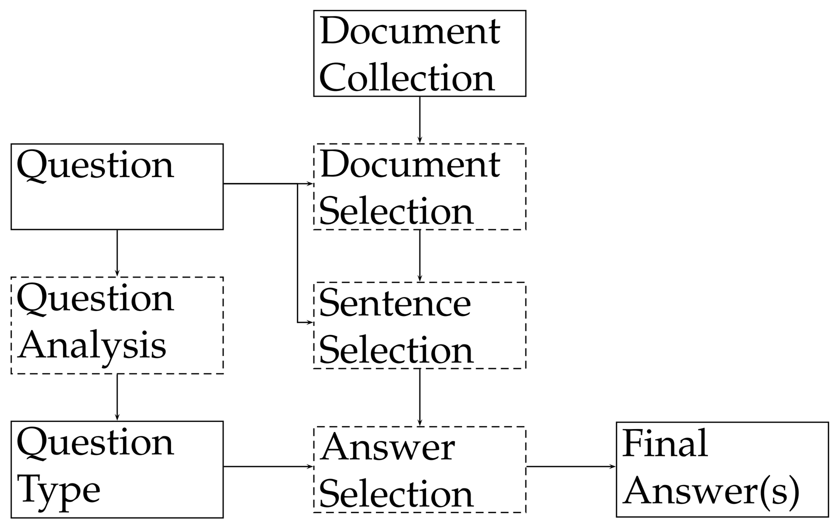
{diego,menno,cassidy}@ics.mq.edu.au

Abstract

Our contribution is centred on a study of Named Entity (NE) recognition on speech transcripts and how it impacts on the accuracy of the final question answering system. AnswerFinder was adapted to the task of question answering on speech transcripts and participated in the QAst pilot track of the CLEF competition. We have ported AFNER, the NE recogniser of AnswerFinder, to the set of answer types expected in the QAst track.

AnswerFinder

AnswerFinder is a question answering system that focuses on shallow semantic representations of questions and text [1, 4]. The underlying idea of AnswerFinder is the use of semantic representations to reduce the impact of paraphrases. The system finds exact answers to questions in large document collections.



AnswerFinder for Speech Transcripts

The special features of speech transcripts make it specially hard to find accurate syntactic and semantic representations of its contents. We therefore simplified AnswerFinder so that it mostly relies on the capabilities of AFNER, the named entity recogniser. The main components of the modified system are:

Named Entity Recogniser : is run over all the documents in a preprocessing step.

Document Selection : returns the full list of documents, unranked.

Question Analysis : uses a decision list of hand-constructed patterns of regular expressions, resulting in question type and set of matching NE types.

Sentence Selection : selects sentences with compatible NEs and ranks them according to word overlap with the question. Uses a simple sentence splitter that relies on punctuation marks and speech turn annotations.

Answer Selection : selects five NEs that are of the expected answer type and have the highest NE scores (duplicate NEs have their scores added). Insert a NIL answer if there are less than 5 answers.

AFNER

AFNER is AnswerFinder's named entity recogniser (NER).

- Specifically designed for the task of QA.
- Aims to increase recall of recognition of entities, at the expense of a possible loss of precision [2, 3].

Features

The features used by AFNER combine various kinds of information:

Regular Expressions	Specific patterns for dates, times, etc
FoundInList	The token is a member of a gazetteer
InitCaps	The first letter is a capital letter
AllCaps	The entire word is capitalised
MixedCaps	The word contains upper case and lower case letters
IsSentEnd	The token is an end of sentence character
InitCapPeriod	Starts with capital letter and ends with period
OneCap	The word is a single capitalised letter
ContainDigit	The word contains a digit
TwoDigits	The word is two digits
FourDigits	The word is four digits
MonthName	The word is a month name
DayOfTheWeek	The word is a day of the week
NumberString	The word is a number word ('one', 'thousand', etc.)
PrepPreceded	The word is preceded by a preposition (in a window of 4 tokens)
PrevClass	The class assigned to the previous token
ProbClass	The probability assigned to a particular class in the previous token
AlwaysCapped	The token is capitalised every time it appears

Process

1. The features are fed to a maximum entropy classifier. Categories in this case are the named entity types prepended with 'B' and 'I' (indicating whether the token is at the beginning or inside a NE respectively), and a general 'OUT' category for tokens not in any entity. So for n named entities, $n * 2 + 1$ categories are used.
2. The resulting classes are merged to build the final entities. We allow nesting in the labels. This way we aim at high recall by allowing multiple interpretations of problematic strings that could be ambiguous.

BPER	ILOC						
IPER	BLOC			BLOC		BDATE	
BLOC	IPER	OUT	OUT	IPER	OUT	IDATE	OUT
Jack	*London*	*lived*	*in*	*Oakland*	*in*	*1885*	.
PERSON				LOCATION		DATE	

Adaptation of AFNER for QAst

We trained the system on a fragment of the BBN corpus (16 documents) and the AMI corpus (26 documents) using NE labels adapted to the QAst task.

Class	Type	# in BBN	# in AMI
ENAMEX	Language	9	0
	Location	2,468	16
	Organization	4,421	27
	Person	2,149	196
	System	0	448
	Color	0	283
	Shape	0	147
	Material	0	267
TIMEX	Date	3,006	9
	Time	96	147
NUMEX	Measure	2,568	293
	Cardinal	0	646

Quality of the Annotations

The quality of the corpus annotations was lower than expected due to various reasons:

Different Application : The AMI annotations were not meant for NER and in many cases they weren't useful. For example, the entity type "Person" would have instances like *industrial designer*.

Poor Quality : The quality of some of the annotations of the AMI corpus was poor. In at least two of the 26 meetings the contents of named entities seemed to be random strings.

Processing Errors : After submitting the results to QAst, we found a bug in our corpus processing script which resulted in some named entities having extra words included in them.

Results

We participated in all the QAst tasks:

CHIL$_M$ Manual transcripts from the CHIL corpus of lectures;

CHIL$_A$ Automated transcripts from the CHIL corpus;

AMI$_M$ Manual transcripts from the AMI corpus of meetings; and

AMI$_A$ Automated transcripts from the AMI corpus.

Each task had two runs:

full Using the complete AFNER.

noML Disabling the ML component and using lists and regular expressions only.

Run	Questions	Correct Answers	MRR	Accuracy
full-CHIL$_M$	98	17.35%	9.98%	6.12%
noML-CHIL$_M$	98	16.33%	9.44%	5.10%
full-CHIL$_A$	98	14.29%	7.16%	3.06%
noML-CHIL$_A$	98	12.24%	5.88%	2.04%
full-AMI$_M$	96	35.42%	24.51%	16.67%
noML-AMI$_M$	96	33.33%	26.39%	20.83%
full-AMI$_A$	93	19.35%	11.24%	6.45%
noML-AMI$_A$	93	22.58%	14.10%	8.60%

There are no statistically significant differences between full and noML versions (T-test on RRs). The small training corpus and the presence of annotation errors in the AMI corpus made the machine learning component of AFNER ineffective.

References

[1] Diego Mollá and Menno van Zaanen. Answerfinder at TREC 2005. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proc. TREC 2005*. NIST, 2006.

[2] Diego Mollá, Menno van Zaanen, and Luiz A.S. Pizzato. Named entity recognition for question answering. In *Proceedings ALTW 2006*, page 8 pages, 2006.

[3] Menno van Zaanen and Diego Mollá. A named entity recogniser for question answering. In *Proceedings PACLING 2007*, 2007.

[4] Menno van Zaanen, Diego Mollá, and Luiz Pizzato. Answerfinder at trec 2006. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings TREC 2006*, page 8 pages, 2007.

Acknowledgements

This work is supported by the Australian Research Council, ARC Discovery grant DP0450750.