

# Designing a Multi-Stream System for Answering Dutch Questions

Valentin Jijkoun Gilad Mishne Maarten de Rijke

Language & Inference Technology Group, U. of Amsterdam

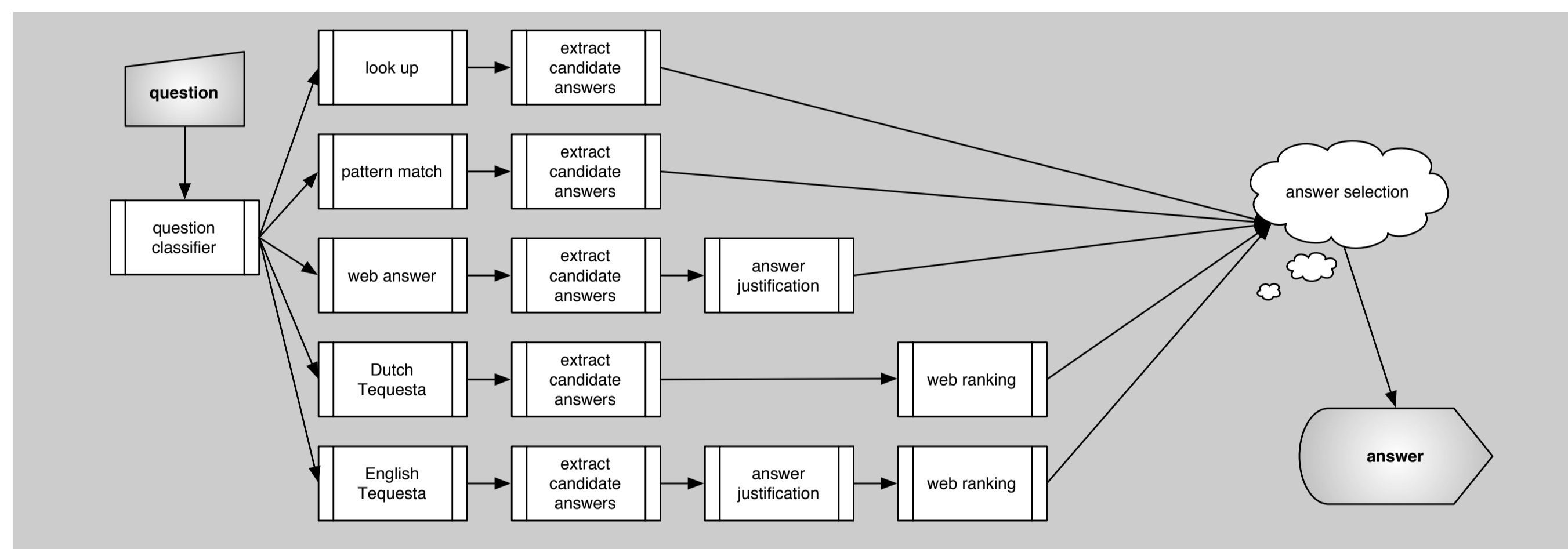
## Introduction

Whereas document retrieval systems aim to find relevant documents in response to keywords provided by a user, question answering systems take questions as input and return answers. In open-domain question answering, the answers are extracted from an unstructured text collection.

So far, most research in question answering has focused on developing systems for English. Our aim was to develop a system for the Dutch language. Given the relative lack of language resources for Dutch, we put a strong emphasis on data-driven methods.

## Architecture

Some answering strategies are beneficial to all question types, and others only for a subset. For this reason we implemented a *multi-stream* system: a system that includes a number of separate and independent subsystems, each of which is a complete standalone QA system that produces ranked answers, but not necessarily for all types of questions. The system's final answer is then taken from the combined pool of candidates.



Our question answering system consists of five independent streams, ranging from linguistically informed to pattern-based to mostly data-driven:

- **Table Lookup:** use pre-constructed specialized knowledge bases
- **Pattern Match:** search for answer patterns generated from a question
- **English Tequesta:** translate questions into English and use Tequesta [3], a linguistically informed QA system for English
- **Dutch Tequesta:** Tequesta with the language-specific components changed to Dutch (named entity and part-of-speech tagging, lemmatization, usage of WordNet, etc.)
- **Web Answer:** transform questions to Web queries, use Google to retrieve snippets of relevant documents from the Web and retain suitable phrases that occur significantly often

While each of the streams in our multi-stream architecture is a question answering system in its own right, they share a number of components:

- **Answer Justification:** given an answer found outside the test corpus, find a supporting document in the collection by means of information retrieval techniques
- **Web Ranking:** use Web hit counts to re-rank and normalize confidence values for answers from different streams
- **Answer Filtering:** filter out strings that are unlikely to answer the question, remove “noise” around answer strings, and merge similar answers
- **Answer Selection:** select the best answer candidates, based on question type and confidence values provided by the streams
- **NE tagger:** statistical n-gram tagger [1] trained on the Spoken Dutch Corpus (CGN) [4] combined with a tagger based on regular expressions

## A Closer Look at Two Streams

Within the *Pattern Match* stream we analyze questions and automatically generate Perl patterns for possible answers. The patterns were developed using a small training set. To identify answers for an input question, the patterns are run against the text collection.

The patterns incorporate shallow morphological information, but they do not use part-of-speech or deeper linguistic information.

- Question In welke Amerikaanse staat ligt San Francisco?
- Pattern `San Francisco\s+ligt\s+in\s+(\S+)`
- Pattern `San Francisco\s+in\s+(\S+)`
- Answer geslaagd (**incorrect**)
- Document text ... de universiteit van Californië in San Francisco in geslaagd om...

- Question Welk land is de grootste olie-producent ter wereld?
- Pattern `([^\?]+)\s+is\s+de\s+grootste\s+producent\s+van\s+olie`
- Pattern `([^\?]+)\s+is\s+de\s+grootste\s+olieproducent`
- Pattern `([^\?]+)\s+is\s+de\s+grootste\s+olie-producent`
- Answer Saoedi-Arabië (**correct**)
- Document text Saoedi-Arabië is de grootste olieproducent ter wereld en...

For our *Table Lookup* stream we mine the text collection (off-line) to extract specific types of information. Some noise reduction is performed, and a cascaded lookup mechanism is used to find information relevant to question terms, allowing for non-exact matches.

An Extract from the <i>Locations</i> Table.			
Location 1	Location 2	Justification	Frequency
Bariloche	in Argentinië	NH19951016-0048	4
Barrow County	in Verenigde Staten	NH19950524-0117	2
Baskenland	in Spanje	NH19951130-0117	2
Basra	in Irak	AD19940104-0065	1
Basra	slechts vier kilometer van de grens met Iran	AD19941015-0007	1
Bathmen	bij Deventer	AD19940708-0171	1
Batna	350 kilometer ten oosten van Algiers	NH19940314-0029	1
Baucau	ten oosten van Dili	AD19950110-0064	1

For table creation and lookup we use shallow morphological information, but no part-of-speech or deeper linguistic information.

- Question Wie is de president van Zuid-Korea?
- Answer Tae Wo (**incorrect**)
- Table entry [Tae Wo; voormalige president van Zuid-Korea; AD19951128-0096; 1]
- Document text Op de dag dat Roh Tae Wo, de voormalige president van Zuid-Korea,...
- Question Waar staat de afkorting WTO voor?
- Answer Wereldhandelsorganisatie (**correct**)
- Table entry [WTO; Wereldhandelsorganisatie; NH19951124-0056; 53]
- Document text ... toetreding tot de Wereldhandelsorganisatie (WTO)...

Tables Used in the System.

Table	Facts extracted	Unique facts	Questions	
			Relevant	Answered (Correct)
Abbreviations	14575	6095	9	8 (8)
Adjective-location	2328	957	14	6 (4)
Locations	4931	4202		
Capitals	1922	465	10	9 (9)
Currencies	41	26	1	1 (1)
Inhabitants	39	38	8	7 (7)
Leaders	10740	2456	30	29 (16)
Roles	9717	8954		
All tables	44293	23193	72	60 (42)

## Results

For testing purposes we used the Dutch CLEF corpus, a collection of newspaper articles from 1994–1995, taken from the Dutch daily newspapers *Algemeen Dagblad* and *NRC Handelsblad* [2]. The total corpus size is about 500MB (72 million words). The set of test questions consists of 200 factoid questions, about 10% of which had no known answer in the corpus.

Lenient Evaluation Results of CLEF 2003 Question Set.

# Questions	Only		Without	All Five Streams
	Table Lookup	Table Lookup	Table Lookup	
200 (all)	54 (27%)	64 (32%)	89 (45%)	
187 (with answer)	41 (22%)	51 (27%)	76 (41%)	

Our error analysis reveals that named extraction recognition and classification is currently an important bottleneck.

## Conclusions & Further Work

Our multi-stream architecture proved to be flexible and effective. It allows us to use multiple strategies to identify different kinds of evidence in support of an answer to a question. Our main findings:

- All streams contribute to the performance of the system.
- The *Table Lookup* stream made a statistically significant difference in the final results, providing correct answers for 58% of all questions relevant to the stream.
- Preliminary experiments suggest that weighted voting between the streams can further improve the overall performance of the system.

In our ongoing and future work we focus on the following issues:

- Informing the answer selection module about past performance of the different streams on different question types.
- Improved named entity recognition and classification for Dutch.
- More linguistically informed lookup mechanisms for the *Table Lookup* stream.
- Incorporation of external information sources (numerous freely available semi-structured knowledge bases) in the *Table Lookup* stream

## Acknowledgments

This research was supported by the Netherlands Organization for Scientific Research (NWO) under project numbers 220-80-001, 612-13-001, 365-20-005, 612.069.006, 612.000.106, and 612.000.207.

## References

- [1] T. Brants. TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied NLP Conference, ANLP-2000*, 2000.
- [2] CLEF: Cross-Language Evaluation Forum. URL: <http://www.clef-campaign.org>.
- [3] C. Monz and M. de Rijke. Tequesta: The University of Amsterdam's Textual Question-Answering System. In *Notebook papers TREC-10*, 2001.
- [4] N. Oostdijk. The Spoken Dutch Corpus: Overview and first evaluation. In *Proceedings LREC 2000*, pages 887–894, 2000.

## Further Information

Please contact [mdr@science.uva.nl](mailto:mdr@science.uva.nl). More information on this and related projects can be obtained at <http://lit.science.uva.nl>. A PDF version of this poster is available at <http://www.science.uva.nl/~mdr/Talks/>.