“How much context do you need?”

An experiment about context size in Interactive Cross-language Question Answering

B. Navarro, L. Moreno-Monteagudo, E. Noguera,
S. Vázquez, F. Llopis and A. Montoyo
Natural Language Research Group (GPLSI)
University of Alicante

Index

1. Introduction.
2. Question translation, disambiguation and expansion.
3. Baseline system: Passage Retrieval system (IR-n) improved with Name Entities Recognition.
4. Experimental system: A Q&A system based on syntactic-semantic similarity.
5. Results.
6. Conclusions and future work.

Introduction

- In order to decide **about the correctness of an answer** shown by an iQ&A system, the sources of information used by an user are:
  - the context in which the (possible) answer appears,
  - (previous) knowledge about the topic,
  - the question itself.
- The **context** is the **main source of information** available for the user.
  - According to the information provided by the context, he/she decides if the answer is the correct one or not, or if it is necessary a refinement of the question.

Introduction

- Problem:
  - The language of the context is different from the language of the query and the language of the user.
  - The user must deal with a language with null or passive knowledge about it.
- Two approaches to solve this problem:
  - to translate the possible answer with its context to the language of the user with a Machine Translation system,
  - or to look for other alternative methods of interaction.
- We are looking for alternative methods of interaction, avoiding the use of Machine Translation systems.

Objective of the experiment

- To know the optimum context size in an interactive cross-language QA framework.
- **Baseline system** shows a complete passage.
  - Maximum context.
  - It has been improved with a named entity recognition system.
- **Experimental system** shows only a clause.
  - Minimum context
  - Pilot version of a Q&A system based on syntactic-semantic similarity.

Objective of the experiment

- Secondary objectives:
  - To know the usefulness of a WSD system based on Relevant Domains applied to question disambiguation.
  - To develop a pilot evaluation of a Q&A system based on syntactic-semantic similarity (experimental system).

Index

1. Introduction.
2. Question translation, disambiguation and expansion.
3. Baseline system: Passage Retrieval system (IR-n) improved with Name Entities Recognition.
4. Experimental system: A QA system based on syntactic-semantic similarity.
5. Results.
6. Conclusions and future work.

Question translation, disambiguation and expansion

- The mother tongue of users is **Spanish**.
  - The questions are written in Spanish
- The text (answers) are written in **English**.
  - Users have passive knowledge of English: they can understand some words/sentences in English, but they can't formulate a question in English correctly.

Question translation

- The questions have been translated to English with three machine translation system available on the web:
  - Systran Babelfish,
  - Reverso Soft.,
  - Google.
- We have selected the common words to two or three translation.
  - If there isn’t any common word between the three translations, we have selected all words obtained.

Question disambiguation

- To obtain the correct sense of each word selected.
- WSD method Relevant Domains.
  - Unsupervised method
  - Relevant Domains are obtained from WN Domains (Magnini & Cavaglia 2000).
    - Domains associated more frequently with a word
  - The system compares context vector and sense vector:
    - Context vector: representative domains of the context words (in the question)
    - Sense vector: domains related with each sense of polysemic words (obtained from the glosses).

Question expansion

- Once we have obtained the correct sense of each word we intend to **expand** each question with a **list of synonyms**.
- We have only one sense per word.
  - The list of synonyms is obtained from WordNet synset.

Index

1. Introduction.
2. Question translation, disambiguation and expansion.
3. Baseline system: Passage Retrieval system (IR-n) improved with Name Entities Recognition.
4. Experimental system: A QA system based on syntactic-semantic similarity.
5. Results.
6. Conclusions and future work.

Baseline system

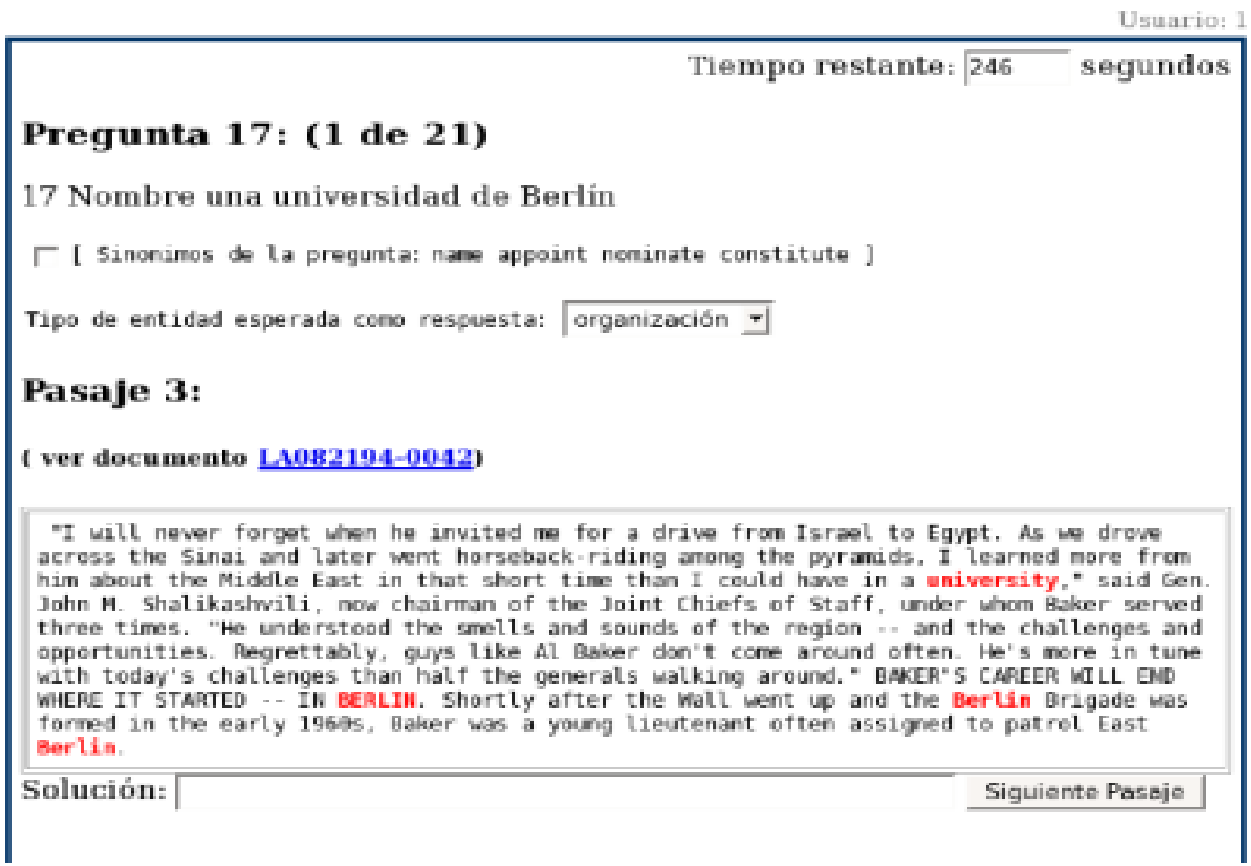
- IR-n system: an passage retrieval system.
  - (See poster)
  - The system shows users the passage (in English) with a possible answer.
  - If the correct answer doesn’t appear in the first passage, the user checks the next one, up to 50 passages.

Baselines system + NE system (I)

- In order to improve the interaction, baseline system uses DRAMNERI, a Named Entity recognition system (Toral et al 2005).
  - Based on rules and gazetteers.
- All entities in the passage similar to the type of entity looking for in the question are shown in different color.
  - Users can change the kind of entity, if it is not correct.
  - Questions words that appears in the passages are shown in different color too.

Baseline system + NE system (II)

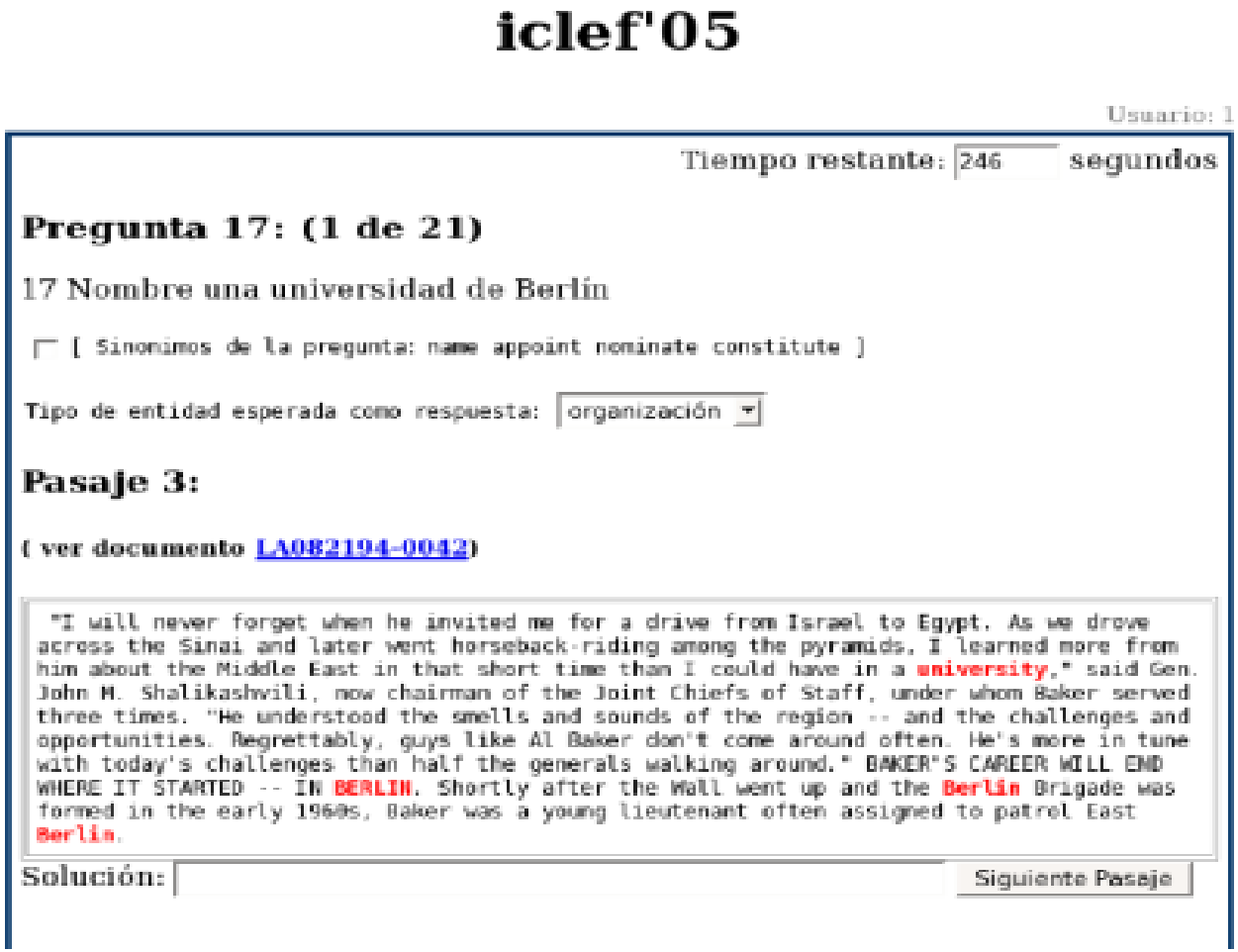
iclef'05



Baselines system + synonyms (I)

- In order to improve the retrieval process, users can **refine the question** with the set of synonyms extracted during question disambiguation.
- In any case, if user want, they can see the whole document.

Baseline system + synonyms (II)



Index

1. Introduction.
2. Question translation, disambiguation and expansion.
3. Baseline system: Passage Retrieval system (IR-n) improved with Name Entities Recognition.
4. Experimental system: A QA system based on syntactic-semantic similarity.
5. Results.
6. Conclusions and future work.

Experimental system

- Pilot QA system based on syntactic-semantic similarity.
  - The system shows users only minimum context: a clause (in English) with the possible answer.
    - Set of words related with a verb in a sentence.
  - If the correct answer doesn’t appear in the first clause, the user checks the next one, up to 250 clauses.

Hypothesis

- Deep semantic relation between a question and its answer.
  - Question is a clause (or more if it is a complex question)
  - Answer appears in a clause.
- **Objective**:
  - To calculate the syntactic-semantic similarity between the question and the clauses in which possible answer appears.

Syntactic-semantic patters

- Both question and possible answers are formally represented as a syntactic-semantic patterns.
- A syntactic-semantic pattern is the subcategorization frame of a verb:
  - A verb: lemma + sense
  - Arguments and adjuncts: head noun (lemma) and it sense(s).
- SS patterns are extracted from passages returned by IR-n.
  - They are processed with a Pos-tagger (Tree-tagger, Schmidt 94) and a syntactic parser (SUPAR, Palomar et al 99)
  - Senses are extracted from EuroWordNet (Vossen 98)

Process

- QA system calculates syntactic semantic similarity between question pattern and all possible answer patterns.
  - The patter with high syntactic semantic similarity with the question represents the clause with the correct answer.

Process. Step 1

- First of all, a filter of proper nouns is applied.
- **Hypothesis**: if a proper noun appears in the question, it must appear in the answer.
  - User needs this information to decide about the correctness of the clause.
- At least, a proper noun of the question must appear in the answer.

Process.

- A syntactic-semantic measure of similarity is applied.

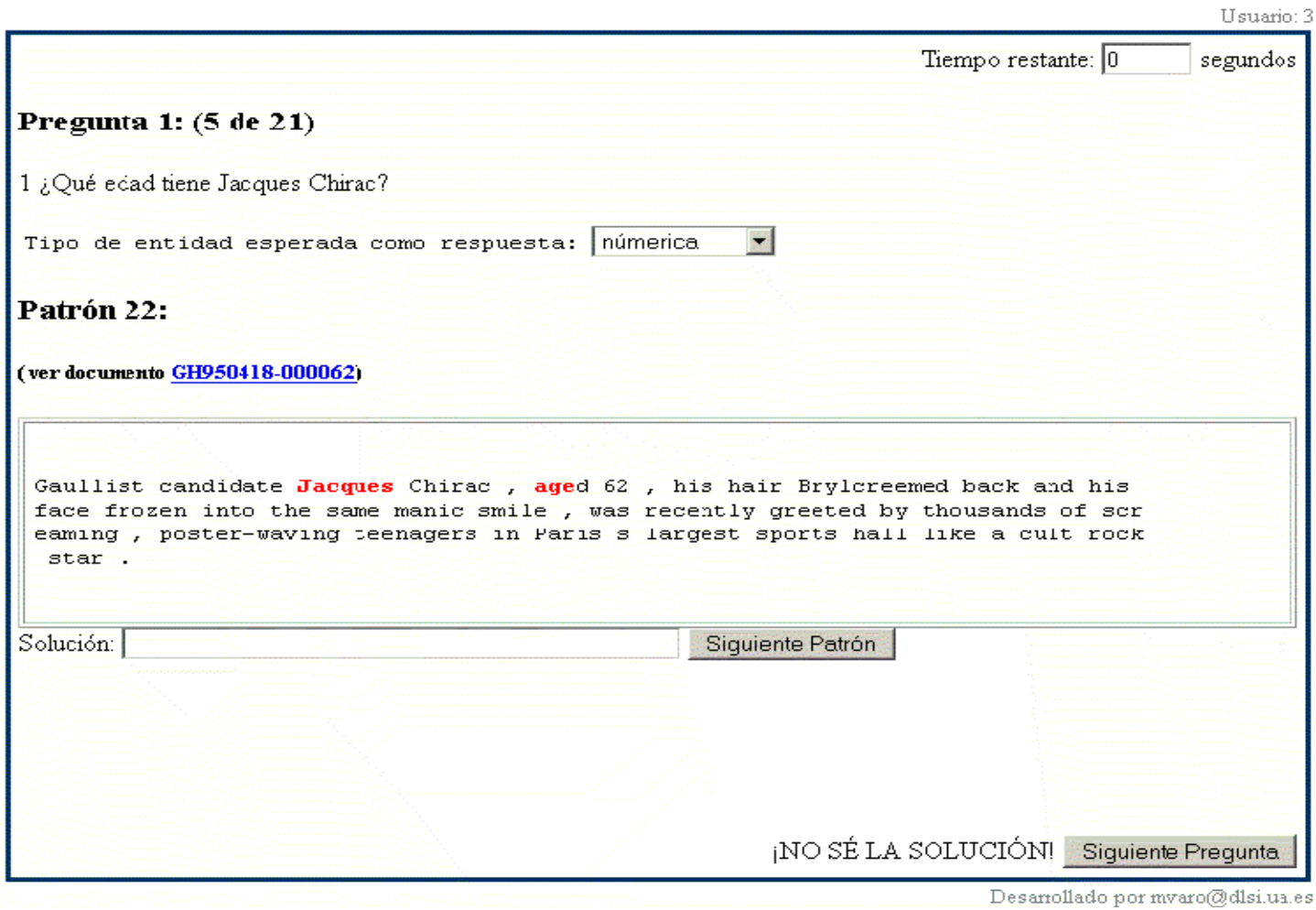
$$Sim(Pq, Pa)=2(SimVpq, Vpa) + (NumAqa+NumPNqa)/2$$

- where:
  - *SimVpq, Vpa* is the semantic similarity between each verb
    - Based on semantic similarity of (Lin 98)
  - *(NumAqa+NumPNqa)* represents the number of equal arguments:
    - Equal lemma of head nouns and equal proper names.

Interaction

- The clauses selected are showed to the user from the most similar to the last one.
- Users must select the clause with the correct answer.

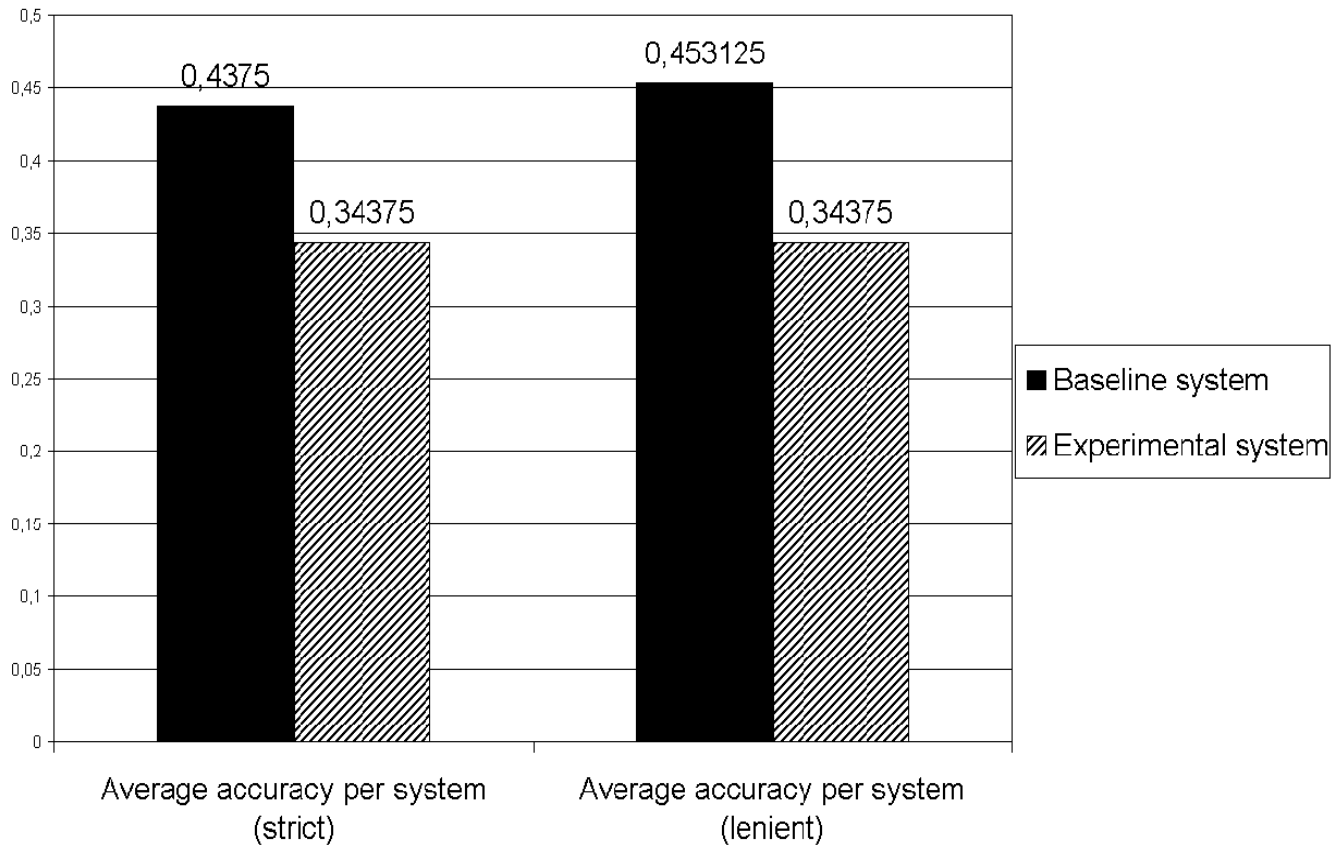
Interaction



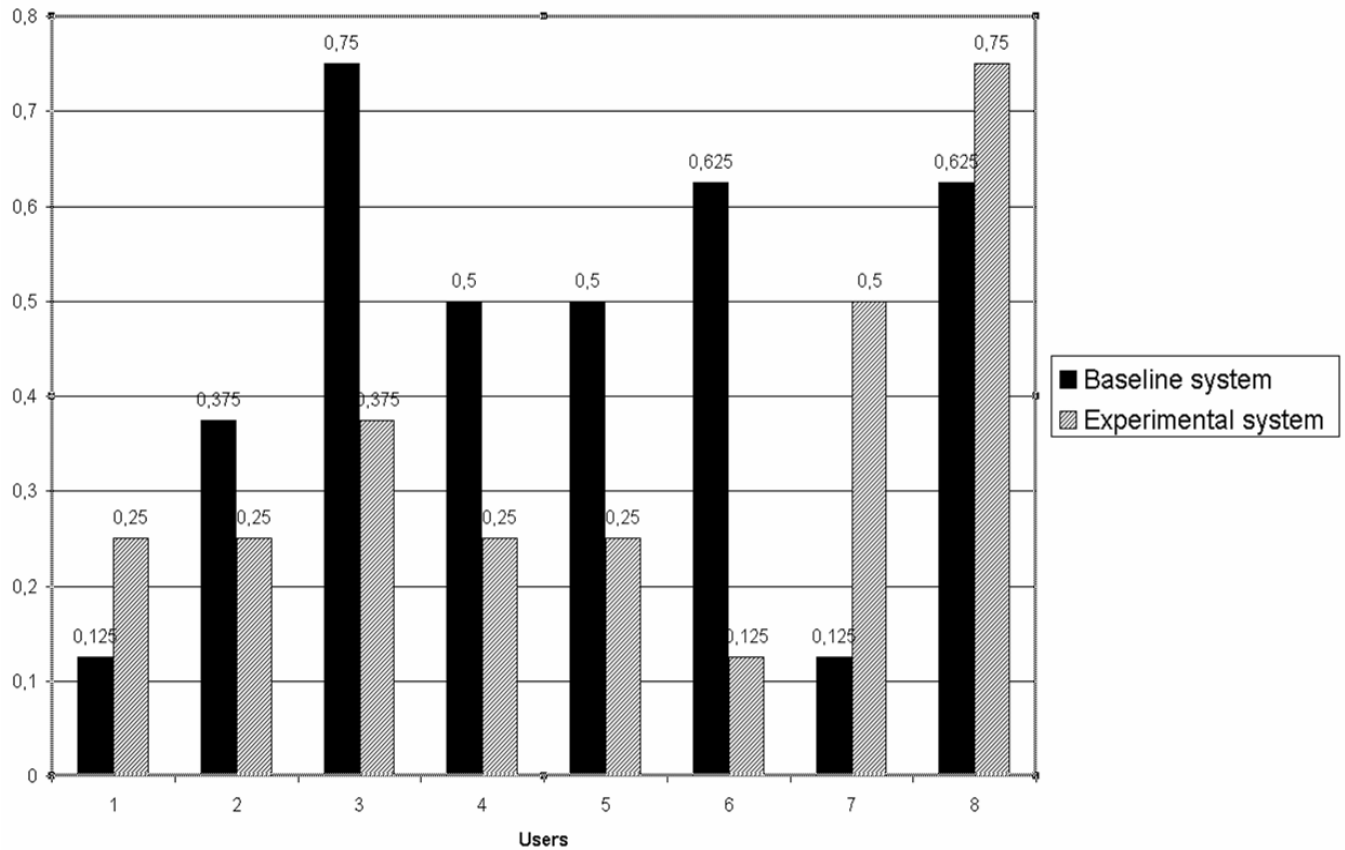
Index

1. Introduction.
2. Question translation, disambiguation and expansion.
3. Baseline system: Passage Retrieval system (IR-n) improved with Name Entities Recognition.
4. Experimental system: A QA system based on syntactic-semantic similarity.
5. Results.
6. Conclusions and future work.

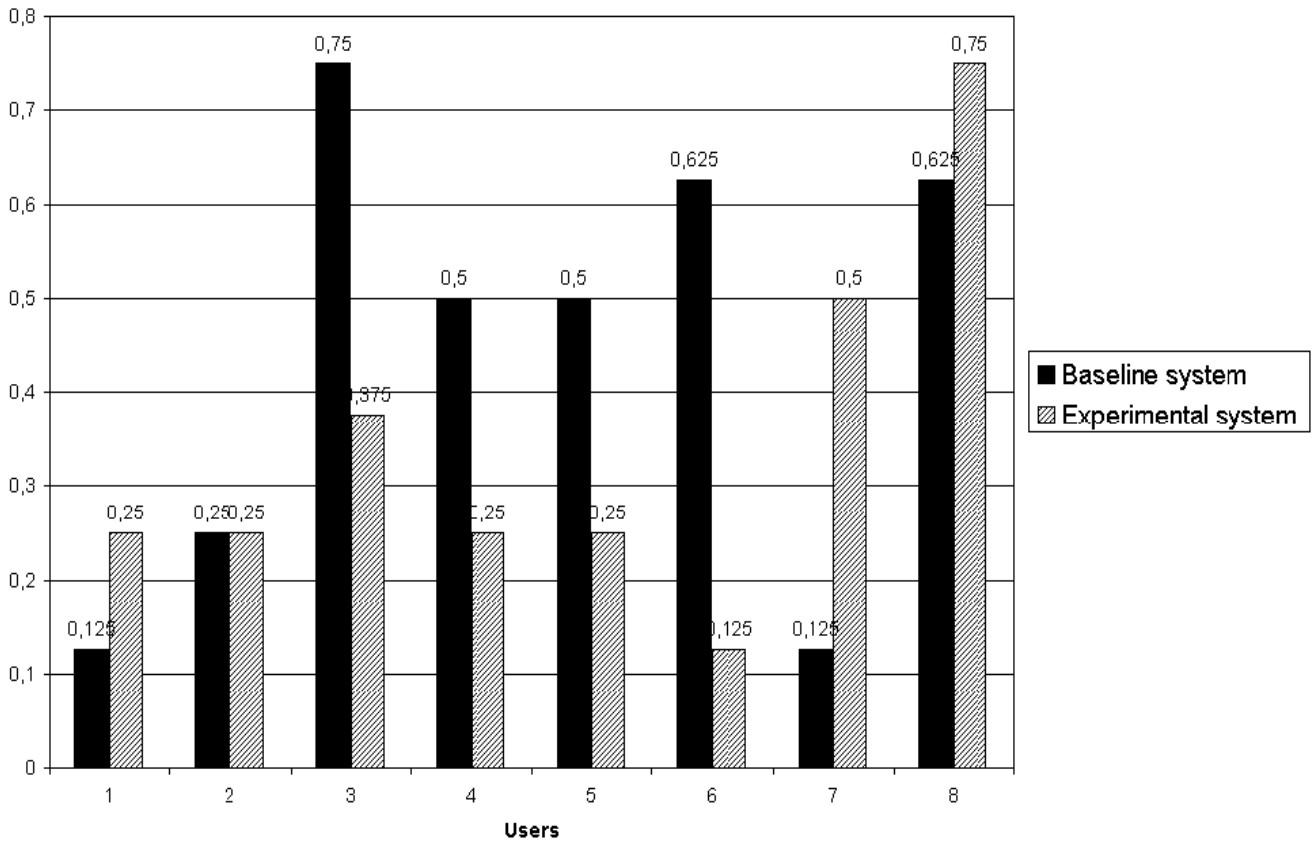
General results



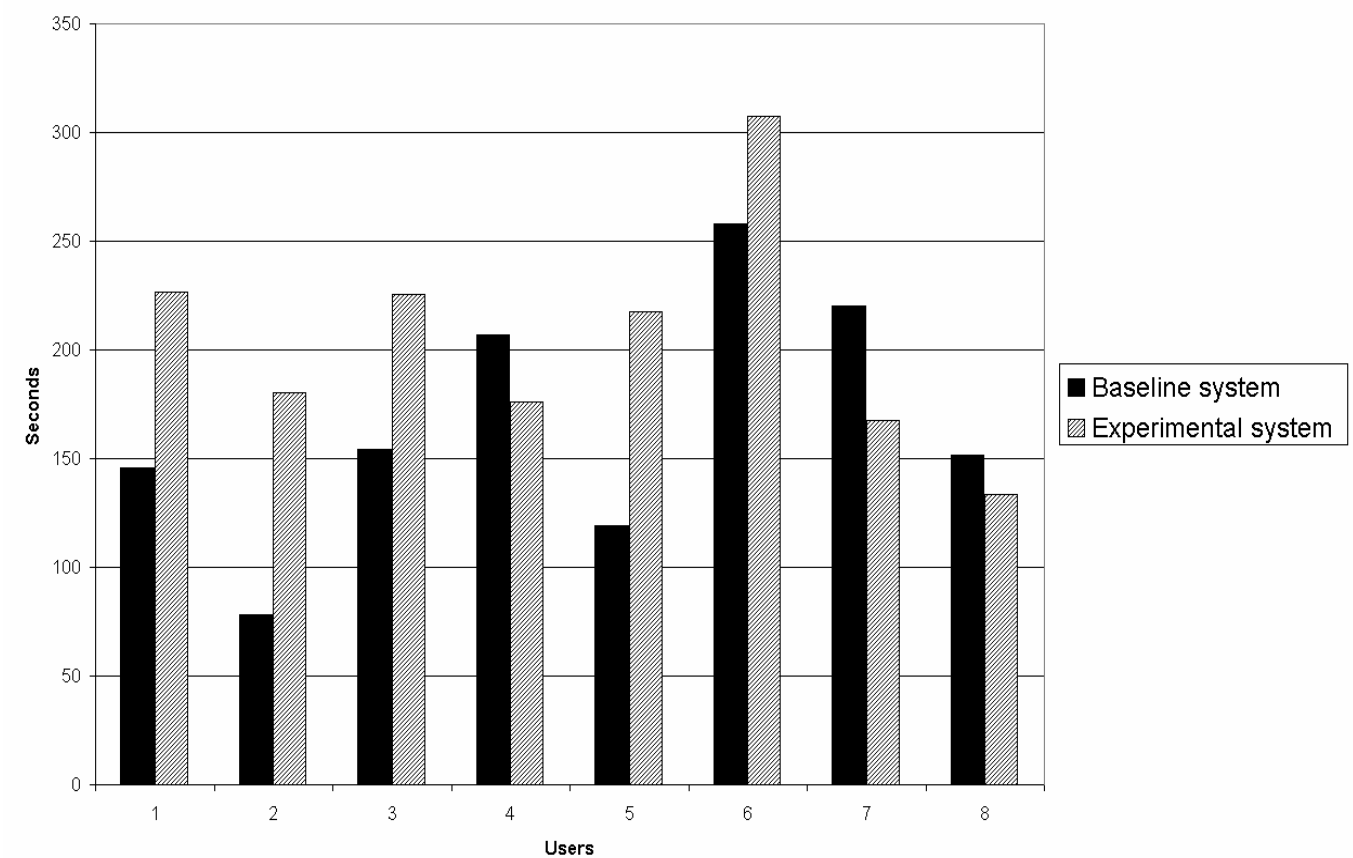
User by user (lenient)



User by user (strict)



Time consuming



NE and Synonyms

- All user said that the information about names entity was useful to locate the correct answer.
- However, users didn’t use synonyms and the expansion of the query.

Index

1. Introduction.
2. Question translation, disambiguation and expansion.
3. Baseline system: Passage Retrieval system (IR-n) improved with Name Entities Recognition.
4. Experimental system: A QA system based on syntactic-semantic similarity.
5. Results.
6. Conclusions and future work.

Conclusions

- It is difficult to establish a fixed context size for an optimum interaction in iQ&A.
- In general, it is better wide context.
- However, for users with poor knowledge of the language of the answer it is more useful and fast interact with sort context.

Future work

- To improve the patter extraction
- To refine the syntactic-semantic measure of similarity.
- To apply semantic parser (semantic roles) in order to detect the correct answer.