A Multilingual Framework for Searching Definitions in Web Snippets

Alejandro Figueroa & Günter Neumann

Language Technology Lab at DFKI

Saarbrücken, Germany

☆ Our interest:

- Developing ML-based strategies for complete end-to-end question answering for different type of questions
  - Exact answers
  - Open-domain
  - Multilingual

☆ Our vision:

- Complex QA system existing of a community of collaborative basic ML-based QA-agents.

- ☆ QA at Trec and Clef evaluation forums have created reasonable amount of freely available corpora
  - Question-Answer pairs
  - Multilingual and different types of questions
  - Contextual information: sentences (mainly news articles)
- ☆ Enables
  - Training, evaluating ML algorithms and
  - Comparisons with other approaches.

☆ Our initial goals:

- Extract exact answers for different types of questions only from web snippets
- Use strong data-driven strategies

☆ Our current results:

- ML-based strategies for factoid, list and definition questions
- Mainly unsupervised statistical-based methods
- Language poor: Stop-word lists and simplistic patterns as main language specific resources
- Promising performance on Trec/Clef data (~ 0.55 MRR)

- ☆ Questions such as:
  - *What is a prism ?*
  - *Who is Ben Hur ?*
  - *What is the BMZ ?*
- ☆ Answering:
  - Extract and collect useful descriptive information (nuggets) for a question's specific topic (definiendum)
  - Provide clusters for different potential senses, e.g., "Jim Clark" → car racer or Netscape founder or …

- ☆ Current SOA approaches:
  - Large corpora of full text documents (fetching problem)
  - Recognition of definition utterances by aligning surface patterns with sentences within full documents (selection problem)
  - Exploitation of additional external concept resources such as encyclopedias, dictionaries (wrapping problem)
  - Do not provide clusters of potential senses (disambiguation problem)
- ☆ Our idea:
  - Extract from Web Snippets only (avoid first three problems)
  - Unsupervised sense disambiguation for clustering (handle fourth problem)
  - Language independent

Why Snippets only?

- ☆ Avoid fetching & processing of full documents
- ☆ Snippets are automatically “anchored” around questions terms → Q-A proximity
- ☆ Considering N-best snippets → redundancy via implicit multi-document approach
- ☆ Via IR query formulation, search engines can be biased to favor snippets from specialized data providers (e.g., Wikipedia) → no specialized wrappers needed
  - Extend the coverage by boosting the number of sources through simple surface patterns
  - Due to the massive redundancy of web, chances of discriminating a paraphrase increase markedly.

☆ Our system's answer in terms of clustered senses:

---------- **Cluster STRANGE** ----------
0<->In epilepsy, the normal pattern of neuronal activity becomes disturbed, causing strange...
---------- **Cluster SEIZURES** ----------
0<->Epilepsy, which is found in the Alaskan malamute, is the occurrence of repeated seizures.
1<->Epilepsy is a disorder characterized by recurring seizures, which are caused by electrical disturbances in the nerve cells in a section of the brain.
2<->Temporal lobe epilepsy is a form of epilepsy, a chronic neurological condition characterized by recurrent seizures.
---------- **Cluster ORGANIZATION** ----------
0<->The Epilepsy Foundation is a national, charitable organization, founded in 1968 as the Epilepsy Foundation of America.
---------- **Cluster NERVOUS** ----------
0<->Epilepsy is an ongoing disorder of the nervous system that produces sudden, intense bursts of electrical activity in the brain.
...

LT-Lab

Example: What is epilepsy?

- Experimental Question Answering System - Mozilla Firefox

bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe

http://www2.dfki.de:8080/experimental-quetal/Foo.do

def 2006

EL ONLINE - Na... LEO Deutsch-Englisch... Anmeldung User-Profil: Günter N... FrontPage - Lucene-h... VFML

I - Experimental Question An... SPIEGEL ONLINE - Nachrichten

DFKI - MULTI LINGUAL WEB QUESTION ANSWERING SYSTEM

Enter Your Question: define:epilepsy Search

DEFINE:EPILEPSY

POTENTIAL SENSE: YEARS

1. EPILEPSY *HAS BEEN A RECOGNIZED UNIQUE DISORDER FOR THOUSANDS OF YEARS.* ...

POTENTIAL SENSE: USUALLY

1. ABSENCE EPILEPSY *THAT CHARACTERIZED BY ABSENCE SEIZURES, USUALLY HAVING ITS ONSET IN CHILDHOOD OR ADOLESCENCE.* ...

POTENTIAL SENSE: UNPROVOKED

1. AND, EPILEPSY *IS A CHRONIC DISORDER, THE HALLMARK OF WHICH IS RECURRENT, UNPROVOKED SEIZURES.* ...

POTENTIAL SENSE: TREATMENT

1. TREATMENT FOR EPILEPSY, *A DISORDER THAT CAUSES SEIZURES*; INCLUDES. ...

POTENTIAL SENSE: RECURRING SEIZURES

1. EPILEPSY *IS A BRAIN DISORDER THAT CAUSES PEOPLE TO HAVE RECURRING SEIZURES.* ...

POTENTIAL SENSE: NERVE

1. EPILEPSY *IS A DISORDER IN WHICH THERE IS EXCESSIVE ELECTRICAL ACTIVITY IN THE NERVE CELLS OF THE BRAIN, WHICH RESULTS IN INVOLUNTARY MOVEMENT OR CHANGES IN AWARENESS.* ...
2. EPILEPSY *IS A BRAIN DISORDER IN WHICH CLUSTERS OF NERVE CELLS, OR NEURONS, IN THE BRAIN SOMETIMES SIGNAL ABNORMALLY.* ...

POTENTIAL SENSE: GROUP

1. IMITATORS OF EPILEPSY *ARE A DIVERSE GROUP THAT INVOLVE CONSIDERATION OF MANY AREAS OF INTERNAL MEDICINE, NEUROLOGY, AND PSYCHIATRY.* ...

POTENTIAL SENSE: FLICKERING

1. PHOTOSENSITIVE EPILEPSY *IS THE NAME GIVEN TO A FORM OF EPILEPSY IN WHICH SEIZURES ARE PROVOKED BY FLICKERING LIGHT ENCOUNTERED IN EVERYDAY LIFE* ...

kann Abwärts Aufwärts Hervorheben Groß-/Kleinschreibung Das Seitenende wurde erreicht, Suche vom Seitenanfang fortgesetzt

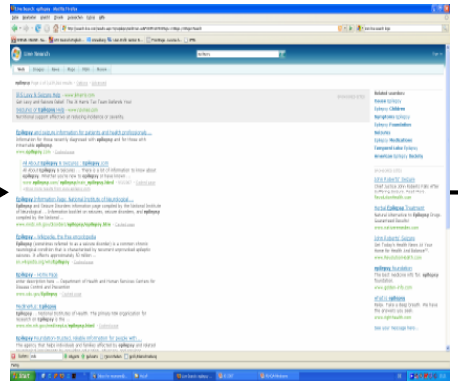
art DFKI - Experimental ... Inbox for neumann@... K12007 WinEdt/MiKTeX - [D:\...] DE

Definition Question

Surface S-patterns

Query

msn live search



Snippts

Surface E-patterns

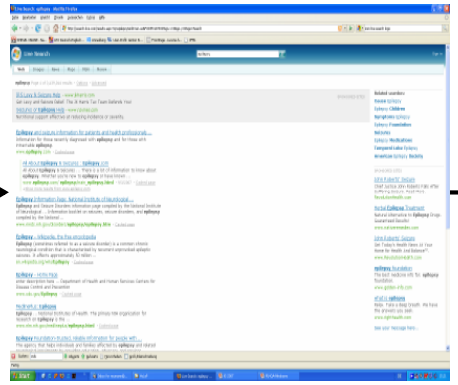
Set of Descriptive Sentences

Definition Extraction

Clusters of Potential Senses

Definition Question

msn live search



Surface S-patterns

Query

Snippts

Surface E-patterns

Set of Descriptive Sentences

Definition Extraction

Clusters of Potential Senses

Seed patterns
- **few**
- **hand-coded**
- **Language-specific**

Seed Patterns

- ☆ Are used to automatically create
  - Search patterns
    → for retrieving candidate snippets
  - Extraction patterns
    → for extracting candidate descriptive sentences from the snippets
- ☆ They are manually encoded surface oriented regular expressions defined for each language
- ☆ Only a few are needed
  - 8 for English, 5 for Spanish

Seed Patterns for English

"X [is|are|has been|have been|was|were] [a|the|an] Y"

"Noam Chomsky is a writer and critical ... "

"[X|Y], [a|an|the] [Y|X] [,|.]"

"The new iPoD, an MP3-Player ,... "

"X [become|became|becomes] Y"

"In 1957, Althea Gibson became the ... "

"X [which|that|who] Y"

"Joe Satriani who was inspired to play ... "

"X [was born] Y"

"Alger Hiss was born in 1904 in USA ... "

"[X|Y], or [Y|X]"

"Sting, or Gordon Matthew Sumner, ... "

"[X|Y][|,][|also|is|are] [called|named|nicknamed|known as] [Y|X]"

"Eric Clapton, nicknamed 'Slowhand'..."

"[X|Y] ([Y|X])"

"The United Nations (UN) … "

Application of Seed Patterns

☆ Some S-patterns for “What is DFKI?”:

- “DFKI is a” OR “DFKI is an” OR “DFKI is the” OR “DFKI are a”…
- “DFKI, or ”.
- “(DFKI)”
- “DFKI becomes” OR “DFKI become” OR “DFKI became”

☆ Some extracted sentences from snippets:

- “**DFKI is the** German Research Center for Artificial Intelligence”.
- “The **DFKI is a** young and dynamic research consortium”
- “Our partner **DFKI is an** example of excellence in this field.”
- “the **DFKI, or** Deutsches Forschungszentrum für Künstliche ... ”
- “German Research Center for Artificial Intelligence (**DFKI** GmbH)”

Extraction of Definition Candidates

- ☆ Approximate string matching for identifying possible paraphrases/ mentioning of question topic in snippets
- ☆ Jaccard measure (cf. W. Cohen, 2003)
  - computes the ratio of common different words to all different words
  - J("The **DFKI**","**DFKI**") = 0.5
  - J("Our partner **DFKI**","**DFKI**") = 0.333
  - J("**DFKI** GmbH","**DFKI**") = 0.5
  - J("His main field of work at **DFKI**","**DFKI**") = 0.1428
- ☆ Avoids the need for additional specific syntax oriented patterns or chunk parsers

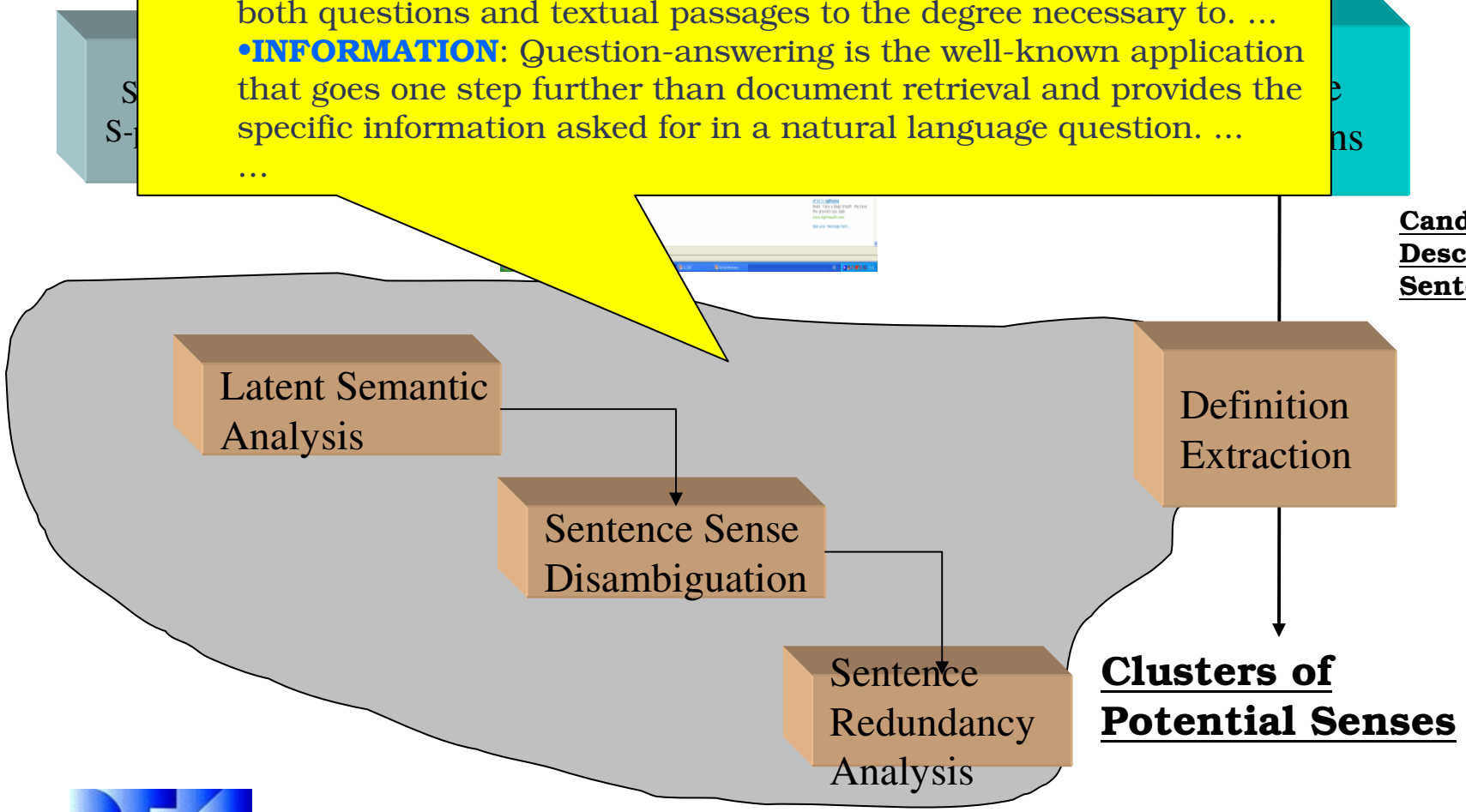
LSA-based clustering into potential senses

- Determine semantically similar words/substrings
- Define different clusters/potential senses on basis of non-membership in sentences

Example: What is Question Answering ?

- **SEARCHING**: Question Answering is a computer-based activity that involves searching large quantities of text and understanding both questions and textual passages to the degree necessary to. …
- **INFORMATION**: Question-answering is the well-known application that goes one step further than document retrieval and provides the specific information asked for in a natural language question. …

…

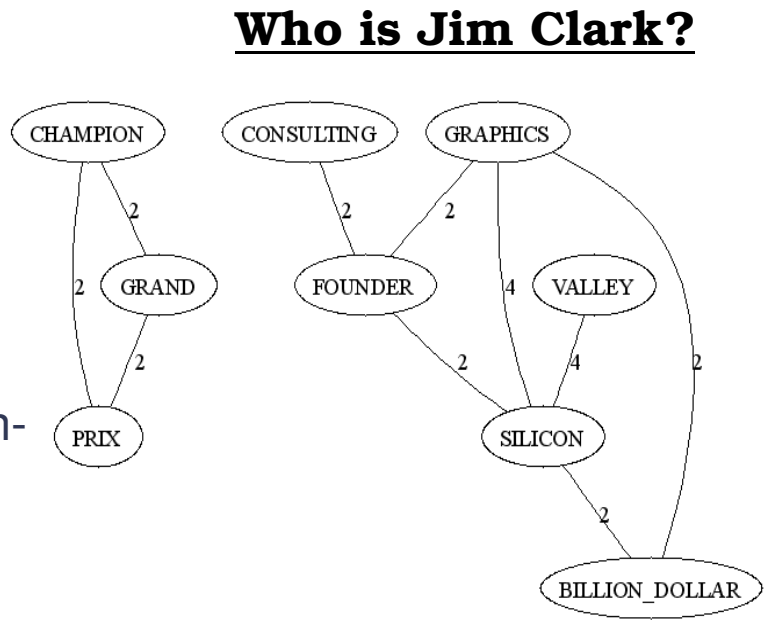


Latent Semantic Analysis

- ☆ Goal: Identify the most relevant terms that semantically discriminate the candidate descriptive sentences.
- ☆ Idea: Use LSA - Latent Semantic Analysis
- ☆ Term-Document matrix construction
  - Document = each candidate sentence + question topic as pseudo sentence ("What is DFKI?" → "DFKI" as pseudo sentence; to dampen possible drawbacks from Jaccard measure)
  - Terms = all possible different N-grams (reduced, e.g., if abc:5 & ab:5 then delete ab:5)
- ☆ Via LSA: select the M (= 40) highest closely related terms to question topic

Clustering

- ☆ Idea: Since words that indicate the same sense co-occur, construct a partition of the descriptive sentences based on the recognition of terms that signal different senses
- ☆ Construct term-term correlation matrix for the M-terms
- ☆ Identify the λ different terms that signal a new sense. Such a sense term:
  - Does not co-occur at sentence level with any already selected sense term
  - Has maximum correlation with the yet non-selected terms
- ☆ Construct λ clusters for the descriptive sentences



Example

- ☆ S_1=John Kennedy was the 35th President of the United States.
- ☆ S_2=John F. Kennedy was the most anti-communist US President.
- ☆ S_3=John Kennedy was a Congregational minister born in Scotland.
- ☆ w_1=35th, w_2=President, w_3=Scotland

term-sentence correlation matrix

$$\Theta = \begin{pmatrix} & S_1 & S_2 & S_3 \\ w_1 & 0 & 0 & 1 \\ w_2 & 1 & 1 & 0 \\ w_3 & 1 & 0 & 0 \end{pmatrix}$$

term-term correlation matrix $\hat{\Theta} = \Theta\bar{\Theta}$

$$\hat{\Theta} = \begin{pmatrix} & w_1 & w_2 & w_3 \\ w_1 & 1 & 0 & 0 \\ w_2 & 0 & 2 & 1 \\ w_3 & 0 & 1 & 1 \end{pmatrix}$$

λ different sense terms: {w_3, w_1} →

Initializing process with randomly selected term, here w_3

Clusters: C_1={S_3} C_2={S_1} C_0={S_2} →

Sentences which do not have a sense term are collected in C_0

NE-readjusted Clusters: C_1={S_3,S_2} C_2={S_1} C_0=∅

Sentences in C_0 with a high NE correlation are reassigned into a corresponding cluster

Removal of Redundancies in Cluster

- ☆ Goal: From each cluster incrementally remove sentences that do not contribute any new information
- ☆ Idea: In each iteration select the sentence for which

$$\max_{S_s \in S_\lambda - \Theta_\lambda} \text{coverage}(s_s) + \text{content}(s_s)$$

Coverage:
Sum of probabilities of those words in S_S which are not already found in previous sentences Θ_λ
→ syntactic novelty

Content:
Sum of the weights of those words in S_S which have a correlation with the question topic (via LSA)
→ semantic bonding

Experiments

- ☆ Two languages: English (EN), Spanish (ES)
- ☆ Baseline algorithm:
  - Query topic using S/E pattern (pattern threshold set to 1 for all)
  - Retrieved snippets S mapped to stream of sentences using JavaRap ("…" as EoS)
  - Remove sentences which have X % word overlap (pair wise check) or which are substrings of other already selected sentences
- ☆ Three different baselines:
  - EN-I: S=300, X=60
  - ES-I: S=420, X=90, patterns from Montes-y-Gomez (Clef 2005)
  - ES-II: S=420, X=90, our patterns

Def-WQA: Results for English

Accuracy of Baseline and MDef-QA for all corpora

Corpus	# Questions	# Answered MDef-WQA/Baseline	# sentences containing nuggets MDef-WQA/Baseline	**Accuracy**
Trec 2001	133	133/81	18.98/7.35	0.94/0.87
Trec 2003	50	50/38	14.14/7.7	0.78/0.74
Clef 2004	86	78/67	13.91/5.47	0.85/0.83
Clef 2005	185	173/160	13.86/11.08	0.89/0.84
Clef 2006	152	136/102	13.13/5.43	0.86/0.85

A set of nuggets as answer of a question

A single nugget as answer of a question

Gold standard

Corpus	F-score (β=5)
Trec 2003	0.52

Trec 2003 best systems (advanced manually developed QA systems on newspaper articles): 0.5 – 0.56

Def-WQA: Results for Spanish

☆ Note that Clef corpora only contain a single nugget (a person or an abbreviation/organization) for a question

Official Clef 2005 systems:
1. 40, 2. 40, 3. 26

Official Clef 2006 systems:
1. 35

Gold standard

Corpora	TQ	ES-I	ES-II	MDef-QA
Clef 2005	50	11	33	32
Clef 2006	42	9	12	22

☆ Problem: Clef corpora consist of news articles from 1994/1995, so data is often outdated in particular for persons

Manual evaluation:
Three human assessors manually checked each descriptive sentence

Manual evaluation

Corpora	TQ	ES-I (AQ/ACCur)	ES-II	MDef-QA
Clef 2005	50	26/0.85	39/0.67	47/0.63
Clef 2006	42	10/0.61	15/0.65	42/0.67

Summary of experiments

- ☆ We achieved competitive results compared to the best Trec and Clef systems
  - We need no predefined window size for nuggets, e.g., Trec uses ~ 125 chars; Clef only person names or abbreviations/organizations
  - MDef-QA computes longer (< 250 chars) but less redundant sentences than the baselines
  - We prefer sentences instead of nuggets for better readability
  - Decrease of accuracy for Spanish prob. due to smaller web space and hence smaller degree of redundancies
- ☆ Problem with a gold standard evaluation:
  - "it is not on my list" → restricted view on recall
  - Inappropriate for Web QA because of "unrestricted" search space

Future work

- ☆ No evaluation of the definition sense desambiguation component so far
  - It seems that it can compute reasonable results, e.g., a good look-and-feel performance
  - But often some senses are distributed across several senses
    - e.g., morphological variations, e.g., for “Akbar the Great” we get senses “emperor” and “empire” because no correlation between the terms
- ☆ Current working focus:
  - Recognition/merging of such distributed senses
  - Explore click behavior of users to adapt clustering (Live QA)
  - Adapting approach to other languages, e.g., German
  - Exploring textual entailment, e.g., for recognizing paraphrases, cf. Wang & Neumann, AAAI-07