Computational Semantics in Information Retrieval

Christof Monz
Institute for Logic, Language and Computation (ILLC)
University of Amsterdam
christof@wins.uva.nl
www.illc.uva.nl/∼christof

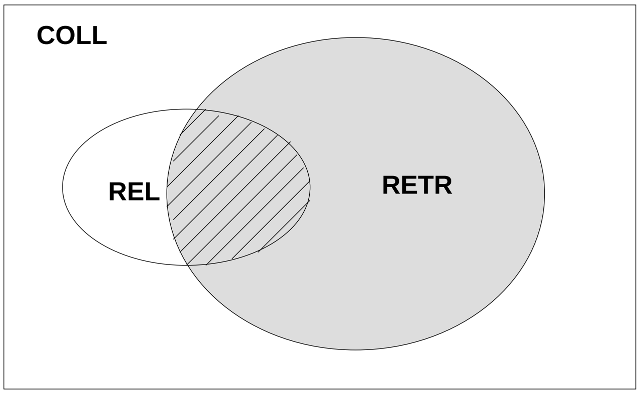
overview

- Information Retrieval in a nutshell
- semantic representation in IR
- inference in IR
- semantic representation in Question Answering
- inference in Question Answering
- conclusions

IR in a nutshell

- goal: fulfill a user's information need
- two kinds of retrieval will be considered:
  - ad-hoc retrieval:
    - present a (ranked) list of relevant documents
    - no user interaction
  - question answering:
    - extract a sentence answering the question from the database
    - generate an answer to the question

basic evaluation of IR systems



- precision: $\frac{|REL \cap RETR|}{|RETR|}$ how many of the retrieved documents are relevant
- recall: $\frac{|REL \cap RETR|}{|REL|}$ how many of the relevant documents are retrieved

semantics and IR

- IR systems represent documents and queries as bags of words
- weights are assigned to words, indicating their importance in general and their importance within a particular document
- the similarity between a document d and a query q is computed as the inner product:

$$\text{sim}(d,q) = \sum_{t \in d \cap q} w_{t,d} \cdot w_{t,q}$$

some shortcomings

- word sense ambiguity: query and document contain the same term t but the occurrences of t differ in meaning (false positive)
- synonyms: although t and t' are different they express the same concept (false negative)
- hyp(er)onyms: t and t' are different but one is a hyponym of the other (false negative)
- phrases: documents where terms occur in the same phrase should be ranked higher
- argument structure: documents where terms occur in similar argument positions should be ranked higher

word sense ambiguity

methodology:

- a word's sense is inferred from its context (i.e. a query or a document)
- words are annotated with their correct sense (e.g. a WordNet synset)
- the similarity between a query and a document is computed on the annotated query and the annotated document

experimental findings:

- word sense disambiguation can lead to a decrease in retrieval effectiveness because the accuracy of most disambiguation algorithms is not very high ($\approx$ 60%)
- mostly positive effects are gained when disambiguation is applied to short queries

synonyms

methodology:

- each query term is annotated with its disambiguated sense
- synonyms are identified as the words that occur in the same WordNet synset
- each query term is expanded by a list of synonyms
- similarity is computed by using the expanded query

experimental findings:

- for long queries, no significant improvement in effectiveness
- for short queries improvements of up to 35% are reported

hyp(er)onyms

methodology:

- each query terms is annotated with its disambiguated sense
- each query term is expanded by a list of hyponyms up to a certain depth
- similarity is computed by using the expanded query

experimental findings:

- similar to expanding with synonyms

phrases

methodology:

- (partial) parsing techniques are used to extract phrases
- alternatively, phrases are identified statistically
- phrases are normalized

experimental findings:

- several results report no difference between statistical and NLP parsing with respect to retrieval effectiveness
- there is no significant improvement in retrieval effectiveness in general, but in several cases there is an improvement of the precision in low recall regions

summing up so far

- word sense disambiguation: significant improvement when queries are short
- synonyms: significant improvement when queries are short
- hyp(er)onyms: significant improvement when queries are short
- phrases: small improvement in low recall regions

in general, the positive effects of applying NLP to IR are very modest and best results are yielded if applied to short queries

question answering

- question answering is more sensitive to the locality of information within a document
- argument structure is essential, contrary to ad-hoc retrieval
- answer to a question can depend on information which is spread over several documents
- the amount of information which is presented to the user is very limited

QA information needs

some example questions from AskJeeves:

- *Are there really such things as shrunken heads?*
- *What movie took the longest to film?*
- *What is mashuganas?*
- *How many of the U.S. Congress and Senate are graduates of Bob Jones University? And what are their names?*

some issues in question answering

- type of question has to be identified
- argument structures within the database have to be identified
- more complex data structures require a softening of the matching criterion
  - argument structures have to be normalized
  - allow synonyms, hyponyms, etc. to match

semantic representation revisited

how to represent argument structures?

- templates
  - argument positions and their values are represented as attribute-value structures
  - inference amounts to (partial) template matching
- description logics
  - argument positions are represented as roles and values as concepts
  - inference amounts to subsumption checking

representation by templates

John Blair was acquired last year by Reliance Capital Inc.

```
agent:     reliance_capital_inc
event:     acquire
patient:   john_blair
time:      last_year
```

drawing inferences on templates

inference $\equiv$ (partial) template matching

problems:

- values themselves can be complex
- matching (just as IR systems mentioned before) have to consider:
  - word sense ambiguity
  - synonyms
  - hyp(er)onyms

description logics

Constructor Name	Syntax	Semantics
concept name	A	$A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$
conjunction	$C \sqcap D$	$C^{\mathcal{I}} \cap D^{\mathcal{I}}$
disjunction ($\mathcal{U}$)	$C \sqcup D$	$C^{\mathcal{I}} \cup D^{\mathcal{I}}$
negation ($\mathcal{C}$)	$\neg C$	$\Delta^{\mathcal{I}} \backslash C^{\mathcal{I}}$
universal quantification	$\forall R.C$	$\{d \mid \forall d' : (d,d') \in R^{\mathcal{I}} \rightarrow d' \in C^{\mathcal{I}}\}$
existential quantification ($\mathcal{E}$)	$\exists R.C$	$\{d \mid \exists d' : (d,d') \in R^{\mathcal{I}} \wedge d' \in C^{\mathcal{I}}\}$
role name	R	$R^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$
role conjunction ($\mathcal{R}$)	$Q \sqcap R$	$Q^{\mathcal{I}} \cap R^{\mathcal{I}}$

representation by description logic formulas

(1) a. Industry sources put the value of the acquisition at $100 million

b. $\exists\texttt{agent}.(\texttt{source} \sqcap \texttt{industry}) \sqcap \exists\texttt{event.put}$

$\sqcap \exists\texttt{patient}.(\texttt{acquisition} \sqcap \exists\texttt{value}.(100,000,000 \sqcap \texttt{dollar}))$

(2) a. John Blair was acquired last year by Reliance Capital Group Inc.

b. $\exists\texttt{agent.reliance\_capital\_group\_inc} \sqcap \exists\texttt{event.acquire}$

$\sqcap \exists\texttt{patient.john\_blair} \sqcap \exists\texttt{time.last\_year}$

automatic generation of the representations

- description logic formulas can be automatically generated from raw text by apploying partial parsing
- additionally lexical information used to identify the argument frames of verbs and to combine this information with the partial parse

inference in description logics

- a document d is about a query q if
  $\vdash \mathtt{d} \sqsubseteq \mathtt{q}$
- a document d answers a query q if it contains an event structure e s.t.
  $\vdash \mathtt{e} \sqsubseteq \mathtt{q}$

where $A \sqsubseteq B$ iff $A^{\mathcal{I}} \subseteq B^{\mathcal{I}}$

description logics in practical IR

- real life IR systems work on hundreds of thousands of documents
  - although there are very efficient DL theorem provers (FaCT, RACE) applying purely DL inference is likely to turn out intractable
- use a hybrid approach where
  - a standard IR system functions as a pre-fetch filtering out the n most likely documents
  - DL theorem proving can then be applied to this restricted set of documents

conclusions

- ☹ although very counter-intuitive, the impact of adding richer semantic representations to IR is very limited
- ☺ on the other hand, if queries are short, as they are in most real-world applications of IR, semantics can improve retrieval significantly
- ☺ for question answering it seems inevitable to employ additional semantic representations that allow to soften the very strict matching criterion
- ?? to which extent deeper semantic representation improves retrieval remains to be shown by applying those techniques to standard collections such as the TREC QA-track