Digging for information WikipediaQAList@wlv at GikiP

Iustin Dornescu

Research Group in Computational Linguistics,
University of Wolverhampton, UK

September 18, 2008

Table of contents

Outline

Why is GikiP interesting?

- there is no snippet that contains the correct answer
- the system has to examine several articles
- stress on link-aware retrieval

Target Collection: Wikipedia

- Plain text is too ambiguous
- Category DAG $\rightarrow$ a folksonomy
- Templates $\rightarrow$ semistructured data
- Interlingual mappings $\rightarrow$ Cross-Language QA
- Wikipedia has information inherent in the link graph

Outline

GikiP Topics

- 15 open list topics
- topic title and topic description
- 3 languages: English, German, Portuguese
- simple geographical concepts: population size, elevation, length, country of birth

GikiP Topics

GP1

Which waterfalls are used in the film "The Last of the Mohicans"?

Name the waterfalls that have been employed in any of the several adaptations of Fenimore Cooper's book "The Last of the Mohicans" to cinema.

- The topic **type** represents the class of articles that could be a correct answer.
- The specific information that the user requires forms the second part of the topic: **the constraint**.

Outline

Select-Filter Paradigm

The idea:

- identify a starting set of documents: candidate answers
- apply filters to remove unlikely candidates

Outline

Select Stage

- high recall initial step
- map the subject NP to a Wikipedia Category
  - "waterfalls" $\longrightarrow$ *Waterfalls*
  - "portuguese rivers" $\longrightarrow$ *Rivers of Portugal*
  - Vienna circle members or visitors $\longrightarrow$ *Vienna Circle*
- use standard tools (parser, wordnet)
- lexicalized rules: "portuguese rivers" $\longrightarrow$ river Portugal portuguese in of "rivers portugal" $\sim$3

Select Stage

- map the NP to a Wikipedia Category
- select the articles directly assigned to it
- fully expand the category tree (D.A.G.) and select **all** the articles
  - the category links are ambiguous: hypernimy,meronimy,etc.
  - the deeper - the less accurate

Outline

Filter Stage

- a filter is a function over sets of documents
  $f : \wp(W_a) \rightarrow \wp(W_a)$
- the initial set of documents can be modified (add,remove) or completely replaced
- a filter extracts information from an article and decides if it is relevant or not
- simple filters could be combined into complex filters
  - the gap between Natural Language and Query Languages
  - Occam's Razor

Filters - Entity

- **Entity filter**: the article must have a link to or mention an entity (GP1,4,6,8,11,12,14,15)
  - GP1: Which waterfalls are used in the film "The Last of the Mohicans"? *The Last of the Mohicans*
  - GP4: Which Swiss cantons border Germany? *Germany*
  - GP8: Suspension bridges in Brazil. *Brazil*

Filters - Attribute

- **Attribute filter**: a certain fact needs to be extracted and the value is analysed
  - population (GP3,GP7,10): "cities with more than 150,000 inhabitants"
  - countryOfBirth/nationality (GP2,9): "born in Germany"
  - elevation (GP6): "higher than 2000 m"
  - length (GP13): "longer than 1000 km"

Results

- Total Answers: **123** in the three languages
- Correct: **93**
- Average precision: **63.2%** (rank 1)
- Score: **15.815** (rank 1)

Analysis

- simple filters applied to the candidate articles
- geographical knowledge:
  - no ontology
  - no reasoning
  - no database(GeoNames, World Fact Book)
  - all these will be added in order to refine the method and deal with more challenging topics
- straightforward cross-language approach
- ambiguity in the link graph is both good and bad
- 3 out of 15 topics had no result