DefScriber: A Hybrid System for Definitional QA

Sasha Blair-Goldensohn, Kathleen R. McKeown, Andrew Hazen Schlaikjer
Department of Computer Science
Columbia University
New York, NY 10027
{sashabg,kathy,hazen}@cs.columbia.edu

Categories and Subject Descriptors

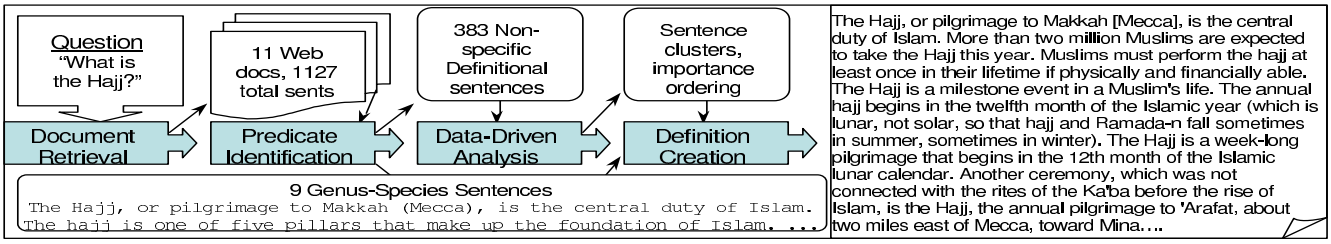
H.3.4 [**Information Storage and Retrieval**]: Systems and Software

Keywords

NLP, Question Answering, Definition, Term

1. OVERVIEW

Much of the effort in Question Answering (QA) has gone into building *short answer* QA systems, which answer questions for which the correct answer is a single word or short phrase. However, there are many questions which are better answered with a longer description or explanation. *Definitional* QA is a developing research area [1] concerned with a subclass of these questions, namely questions of the form "What is X?" DefScriber is a fully implemented system that generates multi-sentence definitions to answer such questions from Internet documents, using an innovative combination of goal-driven and data-driven techniques.



The data-driven techniques in DefScriber use statistically-determined themes in the data to determine content, employing text summarization methods including centroid-based similarity[5] and clustering [4]. The goal-driven part of DefScriber uses a set of *definitional predicates* to identify types of information which should ideally be included in a definition. These predicates (see table) model core properties

Predicate and Description	Instance Example
Genus Conveys conceptual category of the term.	The Hajj is a type of ritual.
Species Describes non-Genus term properties.	The hajj begins in the 12th month of the Islamic year.
Nonspecific Definitional Information relevant to a multi-page definition.	Pilgrims pay substantial tariffs to the rulers of the lands they pass through.

of definitions discussed in the literature [6] and identified in our own research. We use two methods to automatically identify the above predicate types in text: feature-based classification from machine-learned decision trees, and pattern-recognition [3] using patterns manually extracted from a hand-marked corpus. We are currently implementing more predicates, including one that models a term's involvement in cause-effect relationships. A recent evaluation indicated that DefScriber achieves significant improvement over competitive summarization baselines[2]; the below figure traces a test run done for that evaluation.

Our demonstration lets users pose questions interactively via its web interface; robust methods guarantee that a dynamic definition will be generated for any term contained in an Internet document. DefScriber's components are:
Input is a definitional question; desired output length and other parameters may also be specified.
Document Retrieval uses patterns to generate queries based on the definitional question. Queries are sent to a web search engine and documents retrieved.
Predicate Identification searches documents for instances of the definitional predicates Nonspecific Definitional (NSD), Genus and Species. NSD sentences are found and then analyzed for instances of Genus and Species predicates.
Data-Driven Analysis clusters and orders sentences.
Definition Generation combines information from the two previous stages, ordering a Genus-Species sentence first and also applying heuristics over cluster and order information.

Copyright is held by the author/owner.
SIGIR'03, July 28–August 1, 2003, Toronto, Canada.
ACM 1-58113-646-3/03/0007.

2. REFERENCES

[1] ARDA and NIST. *Aquaint R&D Program 12 Month Workshop*, Arlington, VA, 2002.
[2] S. Blair-Goldensohn, K. McKeown, and A. Schlaikjer. A hybrid approach for answering definitional questions. Technical Report CUCS-006-03, Columbia University, 2003.
[3] R. Grishman. *Information Extraction: Techniques and Challenges*. Springer-Verlag, 1997.
[4] E. Hovy and C. Lin. Automated text summarization in SUMMARIST. In *ACL '97 WS on Intelligent Scalable Text Summarization*, pages 18–24, 1997.
[5] D. R. Radev, H. Jing, and M. Budzikowska. Centroid-based summarization of multiple documents. In *ANLP-NAACL WS on Summarization*, 2000.
[6] J. C. Sager and M. L'Homme. A model for definition of concepts. *Terminology*, pages 351–374, 1994.