Chinese News Report and Summarization System

Weiquan LIU
NLP/IR, Intel China Research Center
6F1, North Kerry Center
100020 Beijing, China,
louis.liu@intel.com

Joe F ZHOU
NLP/IR, Intel China Research Center
6F1, North Kerry Center
100020 Beijing, China,
joe.f.zhou@intel.com

Abstract

This demo shows a concept-driven search engine provides human intelligence to the user who desires to seek for relevant information over Internet. Coupled with an auto summarizer and a text-to-speech engine, the retrieved information can be delivered to the user succinctly and in a human conversational manner.

The concept-driven search system consists of a spider, a ConceptNet, and standard information retrieval (IR) engine. The spider is an Internet agent that automatically penetrates into the Cyberspace and gathers information for a specific domain. Represented by a concept network, the ConceptNet is composed by hundreds of nodes, each referring to a specific concept for a chosen domain and pointing to the associations among concepts. The standard IR system serves as the direct interface between the search engine and the user. When the user enters a query, the IR engine returns an initial answer set. Then it pipes the query to the ConceptNet, which checks if there already exists a concept term in the network. If the concept is located, the ConceptNet returns the related concepts and their semantic associations. It is usually the case that the user first gives a general concept as the first query. In return, she gets a broad answer set along with a number of specific concepts from the ConceptNet. Next, the user chooses one (or more) of these specific concepts in order to obtain narrowed and more relevant documents. This process iterates until the user gets targeted result.

Once the user feels satisfies with the returned information, the auto summarizer kicks in. It helps the user locate the main theme of the story and condense it into a short paragraph or in a bulletized format. The summaries can be read or can be listened. When choosing to listen, a smart TTS engine gets initiated. Backed up by an unlimited domain Mandarin TTS system and a large Mandarin speech corpus, the summaries will be spoken out loudly very close to human broadcast.

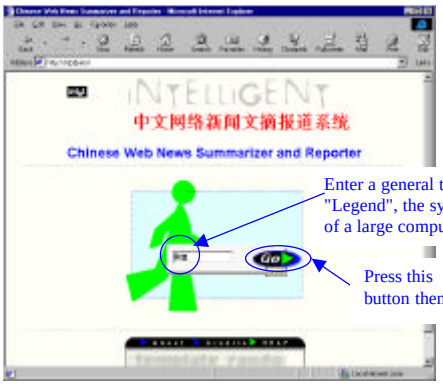
References

Liu W. (1998) *Implementation of a Field Nonspecific Hybrid Automatic Abstracting System.* In: Proceedings of 2nd International Conference on Information Infrastructure. Beijing.

Paice C. D. (1990) *Constructing Literature Abstracts by Computer Techniques and Prospects.* Information Processing and Management 26(1):171-186

Kupiec J. Pedersen J. and Chen F. (1995) *A Trainable Document Summarizer*. In Proceedings of the 18th Annual International ACM *SIGIR* Conference on Research and Development in Information Retrieval.

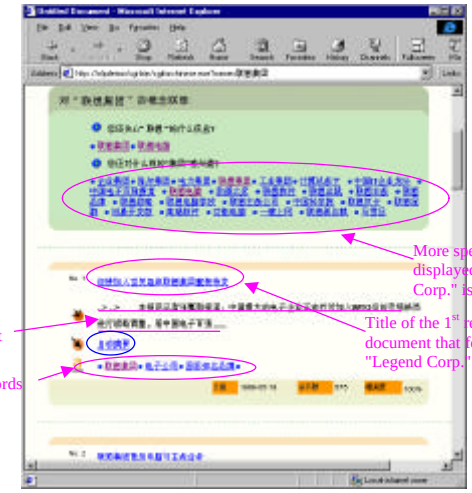
Sample screen shots:



Enter a general term: "Legend", the symbol name of a large computer

Press this button then

"Legend Corp." and "Legend Computer", two more specific terms associated with "Legend" in the concept net



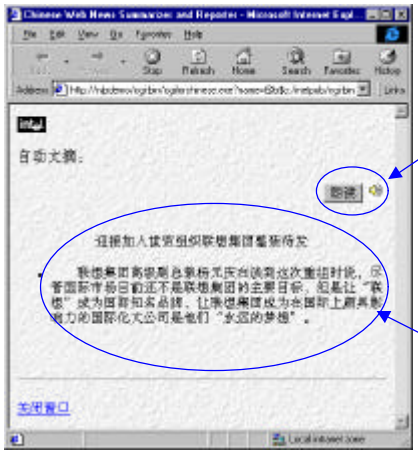
More specific terms are displayed when "Legend Corp." is clicked on

The leading text

Title of the 1st relevant document that found about "Legend Corp."

Keywords

Text summarization



Click on this button to let the TTS engine read the text out.

Single document Summary