1 Introduction

The present paper is a summary of the article *Automated Text Summarisation in SUMMARIST* by Edouard Hovy and Chin-Yew Lin (Mani Maybury, 2001). SUMMARIST is an automated text summarisation system with several modules that together handles the three main tasks: topic identification, text interpretation and text generation. The article outlines the overall system architecture and describes some of its modules.

2 The task of a text summariser

In the introduction, the authors devote a passage to explain their view of what a summariser is, and also requirements posed on such a system. They discuss the distinction between an extract, consisting of portions extracted verbatim from the original text, and an abstract, novel phrasings describing the content of texts. The latter generally requiring stages of topic fusion and text generation.

2.1 Earlier work and different approaches to the field

Previous research in the field that is noted in the paper ranges from simple processes from the late 50’s, looking at privileged positions in the text, lexical cues and locations of topic information, over semantic NLP approaches using representation schemes, frames or templates. The best performing systems today are of the latter type and the authors mention FASTUS (Hobbs,
1992), GE-CMU (Jacobs, 1990) and CIRCUS (Lehnert, 1991) but they also stress the fact that such systems are limited by the fact that frames and templates only can produce an analysis according to what they are predefined to capture. A less semantic and more robust approach is that of employing traditional IR-techniques, scaling down the perspective from a large text collection to a single text. Hovy and Lin note however that traditional pure IR-techniques operate on the word-level and thus miss the concept-level generalisations required for abstract-type summaries. The authors list four main problems with word-level techniques:

- Synonymy (cycle and bicycle both refer to the same vehicle type)
- Polysemy (cycle can mean life cycle or bicycle)
- Phrases (a meaning of a phrase can differ from that of the words in it)
- Term Dependency (terms are not totally independent of each other)

The SUMMARIST-approach is to use IR techniques as far as they are useful and then augment them with symbolic/semantic and statistical methods.

## 3 SUMMARIST

SUMMARIST is still under development and function as a combined research tool and engine to produce summaries (both extracts and abstracts). In each of the three module stages listed below both NLP-processing and symbolic world knowledge (as embodied in WordNet, dictionaries and other resources) is used.

1. **Topic Identification**: subtopic extraction, parameterised to include more or fewer topics or to include only user requested topics. The identification is based on earlier work regarding genre dependent cue words and phrases, text structure, high frequency indicator phrases and discourse structure. Every sentence in the text is assigned a score from each of these separate identification modules and these scores are combined in the internal discourse tree representation to an overall ranking. The module returns the top ranked n% of sentences as its final output.
2. **Topic Interpretation**: merging and fusing of related topic into more general ones (*Mike ate apples, pears and grapes → Mike ate fruit*) and script identification (*she sat down, read the menu and ordered a lobster → she visited the restaurant*). This step is regarded as the most difficult as it requires world knowledge expressed implicitly in the text. To find relations like that in the first example above (*fruit*), SUMMARIST uses a concept generalisation taxonomy built on WordNet concepts. All content words in the text are counted and these numbers are assigned to their respective WordNet-concepts. Then a top-down algorithm picks the most specific generalisation of a set of concepts. The relation in the second example (*restaurant*) is harder to resolve as there is no domain specific knowledge in WordNet. Experiments with full text concept/word clustering and script-based inference, however, have encouraged the authors to proceed with the work trying to retrieve likewise information within smaller parts of a text.

3. **Summary Generation**: There are different levels of summary generation and SUMMARIST will eventually include three output modes: 1) A simple list with extracted keywords or interpreted concepts sorted according to calculated importance. 2) A concatenation of related noun phrases and clauses found by looking at words supporting extracted fuser concepts from the identification module. 3) A reformulation of the extracted material to a dense and nicely phrased new text. For the latter task the authors plan to use an available sentence planner (taking a list with fuser concepts with their most related topics as input) combined with a sentence generator.

Several preprocessing modules are activated before the actual summarisation starts, the modules are: a **tokenizer**, a **part-of-speech tagger**, a **converter** converting tagged text into SUMMARIST representation, a **morpher** finding all token’s root forms, a **phraser** finding WordNet multi-word phrases, a **token frequency counter**, a **weight calculator** calculating and ranking the tf.idf weights of the tokens (multiplication of every term-frequency with the inverse document frequency of that term), as well as a **query relevance calculator** that produces query sensitive summaries based on the number of content words in the user’s query that also appears in that sentence.
4 Conclusion

The SUMMARIST text summarisation system is a running project with in-built system modules in different stages of completion. The knowledge needed to produce abstracts is seldom explicitly included in the text, hence the current construction of a powerful ontology together with a dictionary of topic signatures are important issues. A combination of semantic and statistical techniques are intended to be used to eventually produce three different types of summarisation output, ranging from a simple list of key-words or concepts, over extracted and concatenated text chunks to nicely phrased and reformulated abstracts.

5 Bibliography