Automatic Message Indexing and Full Text Retrieval for a Communication Aid

Stefan Langer
Applied Computing
University of Dundee
Dundee, Scotland
slanger@mic.dundee.ac.uk

Marianne Hickey
Applied Computing
University of Dundee
Dundee, Scotland
mhickey@mic.dundee.ac.uk

Abstract

The aim of the WordKeys project is to enhance a communication aid with techniques based on research in text retrieval, in order to reduce the cognitive load normally associated with retrieving pre-stored messages in augmentative and alternative communication (AAC) systems. In this paper, the differences between traditional information retrieval and the requirements for text retrieval in a communication aid are highlighted. We then present the overall design of the retrieval based communication aid, and describe the morphological analysis module used for indexing and the ranking algorithm in more detail. The system relies on a large lexicon for the automatic indexing of messages and for semantic query expansion. The lexicon is derived from the WordNet database and additionally includes frequency information. Currently, user trials are being carried out to determine the suitability of the approach for AAC.

1 Message retrieval for an AAC system

Currently, there exist different types of communication aids for non-speaking people. Among the systems using natural language, we distinguish two different approaches. The communication strategy can be based on enhanced message composition, or the user can rely on a set of pre-stored messages, together with a selection procedure. It is the latter type of communication aid that will be discussed further here.

A principal deficiency of the current generation of communication aids is the low rate of communication which can be achieved by users. Rates of between 2 and 25 words per minute are typical, which compares poorly to natural speech rates of 150 to 175 words per minute (Foulds, 1980); (Darragh and Witten, 1992). The low communication rate does not encourage either the user of an aid to create messages or a communication partner to maintain attention (Alm et al, 1993). For message selection systems, the low communication rate is partially caused by the fact that many systems rely on retrieval methods that put a high cognitive load on the user. In most systems, the user must remember an access route, or in some cases a code, in order to speak a message. The load placed on the user means that he or she is only able to select from a small number of different things to say.

The reduction of the necessary user input to produce an utterance and the minimization of the cognitive the load on the user in a message-based communication aid can be achieved through efficient message access. A novel approach to reach this is the use of full text retrieval to access a message database. Contrary to most existing message based system, in an AAC system based on text retrieval, in order to select a message, the users do not have to remember any message numbers or another code. They can select a conversational item from the database by entering one or several key words. Appropriate messages will be those containing these words or words related to the key words (Hickey and Page, 1993); (Hickey, 1995).

At a first glance, the implementation of a text retrieval system for AAC users might seem straightforward, as retrieval techniques have been investigated for decades. However, most algorithms suggested in the literature are designed for collections of larger documents, containing several hundreds of words. Little research has been dedicated to the investigation of full text retrieval of short messages such as those used in communication aids. Thus techniques from information retrieval have to be modified con-

siderably to be applicable to the messages communicated by AAC users, which typically contain not more than 20 words. In addition to the difference in length of the messages to be accessed, there is another constraint that affects communication aids to a much higher degree than standard text retrieval systems — the minimal input requirement. In standard text retrieval, queries of 5-10 words are regarded as short queries (Hearst, 1996). This is different for a communication aid. Users of these devices typically have a very low typing rate, and it is desirable that any message from the message database can be retrieved by only one key word, without the need for query refinement.

The state of the art and the named special requirement for a retrieval module in an AAC device suggest the use of enhanced full text retrieval using semantic expansion of queries. A system based on a query expansion technique has the capability of finding messages that contain words that are semantically related to the query words in addition to the messages that contain the query words themselves. Semantic query expansion is especially suited for communication aids, where minimal input and high recall are the key factors. Research in text retrieval has shown that it looks promising to further investigate the use of electronic semantic lexicons both for query expansion and in order to overcome problems of word sense ambiguity (Richardson and Smeaton, 1995). Especially relating to short text, research on image caption retrieval has shown that the recall rate can be considerably higher, if suitable methods of calculating semantic distances between query words and message words are used (Smeaton and Quigley, 1996); (Guglielmo and Rowe, 1996). The measurement of semantic distance can be based on semantic relationship between words. The relationship encoded in many dictionaries and thesauri is synonymy, and often some hypernyms are also included. Both kind of links are relevant for message retrieval. It has been shown that apart from synonyms, which have been used for query expansions for decades, hyponymic links should be considered for text retrieval purposes (Richardson and Smeaton, 1995). The usefulness of hyponymic links has also been evaluated for WordKeys (Langer and Hickey, in preparation). The usefulness of other links, such as meronymy, has yet to be confirmed.

For semantic query expansion through semantically related words, a comprehensive electronic dictionary containing extensive semantic information is needed. Research in electronic lexicography has been very intense during the last years, and many large dictionaries are being built for different languages. Few of those dictionaries, however, are publicly available; and few of those available are suitable for retrieval of unrestricted text. The semantic database WordNet (Miller et al, 1990) has already been successfully used for information retrieval purposes (Richardson and Smeaton, 1995); (Smeaton and Quigley, 1996), and has also been a source for the design of another lexical database for AAC systems, which, like the lexicon used for WordKeys, included additional frequency information (Zickus et al, 1995). The size and coverage of WordNet led to the decision to base the indexing module and the semantic expansion in the WordKeys system on this lexical database.

2 The WordKeys system

WordKeys is a system based on full text retrieval of pre-stored messages. It is typically used in two different settings:

- When the user wants to prepare a communication, new messages are typed in. These messages are automatically indexed and integrated in the system's database.
- In communication mode, WordKeys displays the search field, where the user can type in search words, the list of predicted input words, the list of messages found and the field containing the selected message.

Figure 1 demonstrates the overall architecture of the WordKeys system.

WordKeys is implemented in C++. There is strong emphasis of re-usability of the software, especially the lexicon modules, for other AAC-systems. We have also taken care to provide the possibility of porting the system to languages other than English. The different lexicons are text files and correspond to a simple and clearly specified format. They can be exchanged for lexicons in other languages.

3 Indexing, morphological analysis and message ranking

3.1 Indexing

The WordKeys system offers the possibility of importing any text file to add it to the message database. Additionally, at any stage of a conversation, the user can add a message to the database or modify an existing message. When a message is added to the database, the following actions are performed:

- Tokenization;

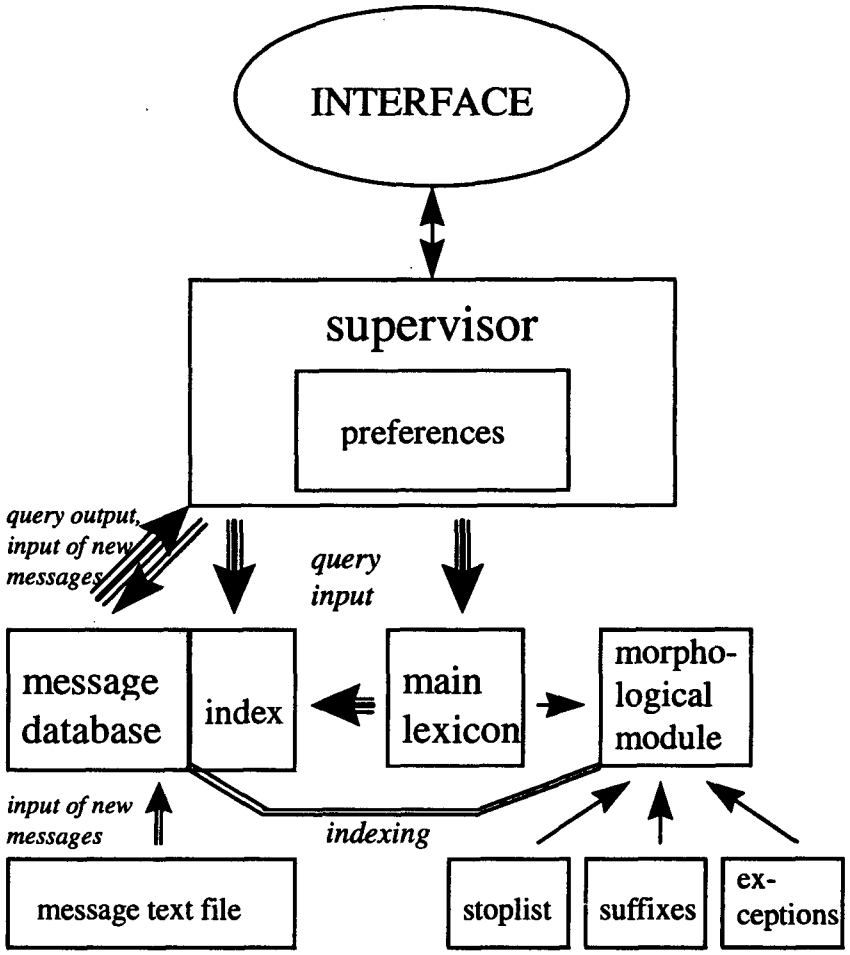

Figure 1: Overall organisation of the WordKeys system

- Morphological analysis: word forms are analysed to find lemmas and roots and to determine their syntactic category;
- The resulting words are looked up in the semantic lexicon to find frequent hypernyms which are added to the list of index words;
- The message with the list of index words is added to the database and its index.

3.2 Morphological analysis

Morphological analysers are available in the public domain. However, we decided to use a custom programmed morphological module, because the output of the available analysers did not correspond to our needs, and, at least for English, a simple analysis is relatively easy to implement. The data used for analysis is partially based on the WordNet morphological information. The morphological module uses an affix list in combination with an exception list and the information about syntactic categories from WordNet. The analysis of a word form is carried out in two steps:

- lemmatization;
- determination of the derivational root (only for semantically transparent derivation affixes).

Lemmatization is lexicon-based. After the affix removal, the unaffixed form is looked up in the lexicon, considering the possible syntactic category returned by the affix removal process. Only if the form is found there, it is accepted as a lemma and added to the message index. Word forms leading

to several possible lemmas are currently not disambiguated. Apart from the lack of disambiguation, we achieved an error-free lemmatisation of all occurring word forms for a trial message database of about 1200 words.

After the lemmatization procedure, a derivational analysis is carried out on the lemmatized word forms. We separate the two steps in order to be able to give the link between a word form and the lemma a higher weight in message access than links between morphologically complex words and their roots. The procedure of distinguishing between the results of inflectional and derivational analysis is consistent with the findings reported in Hull (1996). He concludes that complex stemming algorithms can be slightly more effective than simple ones, and that the removal of derivational affixes is not always desirable. This is especially true for a system such as WordKeys, which uses semantic relationship for retrieval and performs message ranking, which can increase the impact of inaccuracies in the morphological analysis. Semantic relations between a lemma and some word form on the one hand differs considerably from the semantic relations between derived words and their root.

To be able to determine semantically related words without loss of precision, information from the morphological analysis is also used to determine the morpho-syntactic categories of word forms and lemmas. The category can be clearly determined in the following cases:

- a word has one single entry in the main lexicon, which means the word is already a lemma;
- a word form has an inflectional or derivational affix which only occurs with bases of one single morpho-syntactic category.

Removing ambiguities concerning syntactic categories has a certain impact on the performance of the semantic expansion module. The less words with inappropriate syntactic categories are included in the index, the higher precision will be achieved by the system, because less expansions will be generated. For many word forms in the messages, however, the category remains ambiguous. Currently, we are investigating the use stochastic taggers and local grammars for determining syntactic information in these cases.

3.3 Message ranking

When the user has typed in one or several key words and decides to start the search the following tasks are carried out:

- Tokenization: the content of the input field on the interface is parsed into word forms.
- Lemmatization: word forms are analysed to be able to look them up in the lexicon.
- The word forms and lemmas are looked up in the message index. If they are found, the corresponding message numbers are added to the list of retrieved messages.
- The lemmas are looked up in the semantic lexicon to retrieve related words. The relations used for query expansion are dependent on the semantic paths defined in the settings. The related words are re-applied for another query to the index of the message database.

The messages which have been found are displayed on the screen, the order corresponds to their score.

Trials with a number of different settings for the message retrieval algorithm have been carried out to improve message ranking. The ranking algorithm assures that messages which are retrieved, but are not considered very relevant for a query, are put lower in the list or excluded from the display. Conforming to the results of the trials, messages retrieved from the database are ranked according to the criterion of semantic distance between key word and index word. Semantic distance is zero in the beginning of the following list and increases:

- same word form;
- different word form from the key word form (*cars* — *car*);
- other derivation of the root of the key word (*investigation* — *investigate*);
- synonyms of key word (*car* — *automobile*);
- other related words: the semantic paths and their weighting are defined in the settings file. A path is the concatenation of semantic links that are used to get from the input key word to the index word.

Table 1 gives the figures for the message ranking criteria applied in the case of one single key word. For several key words, a combination of the semantic distances for different key words is used for ranking. When several key words are typed in, the message retrieval algorithms is working with an OR- link between search words. However, any message being retrieved by more than one of the key words will be given an increased score; the more key words a message is related to, the better its score.

Description	Weight decrease	Comment
Word in message is same word form as input word	0	exact match, best rating
Word in message is lemmatized in index and matches input word	1	lemmatization leads to less semantic distance than derivational analysis
Word in message is reduced to root in index to match input word	2	derivational analysis
Semantically related word is looked up in lexicon	≥ 5	depends on semantic relation

Table 1: Value determination for message ranking

We will illustrate the message ranking with an example. The messages retrieved from an experimental database for the item *swim* are (in that order):

(1) *Would you like to go for a swim?*

(2) *Normally I don't like swimming, but this Sunday it was so hot that I spent the whole day on the beach and in the water.*

(3) *I'm not a very good swimmer.*

(4) *Shall we go for a dip?*

The first message contains the key word itself; message (2) contains another word form of the same lemma. The third message in the list contains a derivation of the key word. Finally, message (4) is an example of retrieval through semantic query expansion. It contains a synonym (*dip*) of the key word.

3.4 The lexicon for query expansion

One purpose of the main lexicon in WordKeys is to serve as a lexical database for the indexing module when performing morphological analysis. The main function of this lexicon, however, is to serve as a basis for the semantic query expansion. To choose the right lexicon, we had to bear in mind that WordKeys is a retrieval system for unrestricted text. This implies that the system is able to retrieve messages containing any word of the English language apart from extremely domain specific vocabulary.

We decided to use the semantic database WordNet for the following reasons:

- it is very comprehensive;
- it contains most relevant semantic links;
- the information contained in WordNet is stored in text files, and can be easily converted to any other format.

In order to use the information in WordNet for our text retrieval algorithm, some preparation was needed.

- WordNet was converted to a format suitable for the WordKeys software. We chose a format which was easily portable: a text file containing lemmas together with their syntactic category and related words corresponding to the different senses;
- The semantic paths that the WordKeys software uses for query expansion were defined. A semantic path is a series of semantic relations which can be used to reach a lemmatised message word from a lemmatised input key word. This also involved defining weights for the links in order to rank retrieved messages. For example, messages containing synonyms of key words receive a high rating, those containing hypernyms are assigned a lower rating.

Additionally we included statistics over word frequencies in the main lexicon, in order to be able to retrieve hypernyms of words that are useful as index words - these are not necessarily the closest superordinated words in the WordNet hierarchy, but often words occurring several levels higher.

Consequently, in each lexicon entry the following information is stored:

- Syntactic category of word, which is used for morphological analysis and semantic links.
- Frequency (0 if the word is not included in the frequency list). The frequency stored is retrieved from a large database of mainly written text, the British National Corpus (BNC). The list contains the most frequent 8000 words in this corpus; evaluation of a comparison between a frequency counting lexicon and a lexicon without word frequencies are summarized in the next section.

- Links to other words in the lexicon, and specification of the type of link (synonym, hyponym etc.).

4 Evaluation

Formal evaluation of the performance of the semantic retrieval modules is reported in Langer and Hickey (in preparation). The purpose of the evaluation was to look at the benefits of semantic expansion in terms of retrieval success. These trials have shown, that the semantic expansion enhances recall rate considerably without having a major effect on precision. A further improvement was achieved by including frequency information in the lexicon. The evaluation results for the frequency counting lexicon were considerably better than those for the first lexicon derived from WordNet without including frequencies. The frequency counts can be used to determine frequent hyponyms and hyperonyms of index terms, even if they are rather distant in the hyponymic hierarchy, because the method allows the elimination of intermediate level terms that are not interesting for query expansion. Taking into account this information resulted in a higher recall rate, without a major detoriation in precision. For a small, experimental database, 80% of the target messages were correctly retrieved (compared with about 70% with the first lexicon and 50% without query expansion).

Currently we have started the second evaluation phase. A non-speaking person who has already had experience with message based systems for several years, and who uses a relatively large pool of pre-stored messages, is evaluating WordKeys in real communication settings. The main purpose of this study is to investigate the following points:

- How high is the recall rate in real communication settings?
- How useful is the semantic expansion module, i.e. how often does it play a crucial role for successful retrieval?

Apart from these more technical points, the interface and ease of use will also be a evaluated.

5 Conclusions

We have detailed the reasons which lead to the design of a communication aid for non-speakers based on ideas from text retrieval with semantic expansion, and we demonstrated the overall design of the prototype. The main differences between standard information retrieval and text retrieval for an AAC system were presented, namely the size and type of texts retrieved by the system and the necessity to minimize the cognitive load, which leads to the minimal input requirement. In the detailed system description, we have shown that a precise morphological analysis can be achieved — at least for English — with relatively low effort, if we use data from publicly available resources. The morphological module is indispensable to be able to enhance the system with a query expansion algorithm, which is needed to satisfy the minimal input requirement for communication aids.

Several ideas for improving the text retrieval algorithms and WordKeys and their inclusion in other communication aids are still waiting to be realised. One idea is to use a semantic lexicon that is able to learn from the input. This would mean that successful semantic links will get a higher weight than other ones. New links could be added based on knowledge of the user's message selections. Interactions where initial search words do not retrieve relevant messages could be recorded.

A further aim is the integration of the retrieval module with other AAC systems. WordKeys is not designed to assist all kinds of communication. Integration with other AAC software should be investigated, such as software designed for unique text entry (word prediction systems) and for the rapid use of quick conversational fillers.

Acknowledgements

The current phase of the WordKeys project is funded through a European HCM/TMR fellowship for 20 months (January 1996 - September 1997). Evaluation equipment has been purchased with a donation from the Anonymous Charitable Trust.

References

Alm N., Murray I.R., Arnott J. and Newell A.F. 1993. Pragmatics and affect in a communication system for non-speakers. *Journal of the American Voice I/O Society. Special Issue: People with disabilities*, March 1993, pp. 1-15.

Darragh J. and Witten I. 1992. The Reactive Keyboard. Cambridge. Cambridge University Press.

Foulds R. 1980. Communication rates for non-speech expression as a function of manual tasks and linguistic constraints. In *Proceedings of the International Conference on Rehabilitation Engineering*, Toronto, RESNA, pp. 83-87.

Guglielmo E.J. and Rowe N.C. 1996. Natural-language retrieval of images based on descriptive

captions. *ACM Transactions on Information Systems*, 14, 3 (July 1996), pp. 237-267.

Hearst M. 1996. Improving Full-Text Precision on Short Queries using Simple Constraints. In *Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval (SDAIR)*, Las Vegas, April 1996.

Hickey M. 1995. Communication enhancement in an aid for severely disabled people. PhD Thesis. Coventry University.

Hickey M. and Page C. J. 1993. Polyvox: Flexible message selection in a communication prosthesis for non-speakers. In *Proceedings of the 2nd European Conference on the Advancement of Rehabilitation Technology*, May 26 - 28, 1993, Stockholm, Sweden, Section 11.3, pp. 89 - 91.

Hull D. 1996. Stemming Algorithms: A Case Study for Detailed Evaluation. *Journal of the American Society for Information Science*, vol 47, Number 1 (January), pp 70-84.

Langer S. and Hickey M. in preparation. Using Semantic Lexicons for Intelligent Message Retrieval in a Communication Aid. Submitted to *Journal of Natural Language Engineering, special issue on Natural Language Processing for Communication Aids.*

Miller G. A. et al. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4).

Richardson R. and Smeaton A. 1995. Using WordNet in a Knowledge-Based Approach to Information Retrieval. Working paper CA-0395, School of Computer Applications, Trinity College Dublin.

Smeaton A. and Quigley I. 1996. Experiments on Using Semantic Distances between Words in Image Caption Retrieval. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval*, Zürich, pp. 174-180.

Zickus, W. M., K. F. McCoy, P. W. Demasco and Pennington C.A. 1995. A lexical database for intelligent AAC systems. In *Proceedings of the RESNA '95 Annual Conference*, pages 124–126, Arlington, VA. RESNA Press.