Integrated Language Technologies for Multilingual Information Services in the MEMPHIS Project

Walter Kasper, Jörg Steffen, Jakub Piskorski, Paul Buitelaar

German Research Center for Artificial Intelligence
Stuhlsatzenhausweg 3
D-66123 Saarbrücken
Germany
{kasper,steffen,piskorsk,paulb}@dfki.de

Abstract

The MEMPHIS project integrates a large set of NLP technologies. An overview of components, their underlying technologies and resources will be presented: language identification, document classification, linguistic analysis, summarization, information extraction, machine translation, knowledge management and crosslingual retrieval.

1. Introduction

The MEMPHIS project (http://www.ist-memphis.org) aims at developing a platform for cross-lingual premium content services, targeting mainly portable thin clients, like mobile phones or PDAs. The core of the system is a cross-lingual transformation layer integrating multilingual information extraction and summarization of source documents, translation to the customers' target languages and knowledge management for extracted information based on an application's domain ontology. These functionalities are based on a rich set of linguistic resources, many of them application and domain dependent. These resources include e.g. statistical models, specific corpora, terminologies, ontologies and extraction grammars. This implies that a corresponding rich toolkit for creating and updating such resources for specific services is indispensable. As proof of concept two applications are developed within the project, differing considerably with respect to their domains, their information sources and service requirements: MediAlert providing information about new media, especially books, on specific themes, and FinAlert providing business information based on news, e.g. merger and acquisition activities. The languages involved are English, German and Italian.

We will present an overview of the major components, resources and technologies to meet the challenges:

- Language Identification and Verification
- Document Classification
- Shallow Linguistic Analysis
- Summarization and Keyword Extraction
- Information Extraction
- Machine Translation
- Knowledge Management and Cross-lingual Retrieval

All levels of processing communicate through XML annotations on the source documents. These annotated documents are stored as analysis documents. The major components are tied together in a flexible architecture which does not predefine a specific flow of information in the system. All components are visible to each other through a registry, allowing each component to request others if it needs information from it.

2. Language Identification

Since MEMPHIS retrieves documents in various languages and the documents usually do not specify the language, the first processing step consists in identifying the document language. We use a robust statistical approach based on Markov models using character level n-grams (Dunning, 1994). For each language supported, it requires only between 1500 and 2000 words to train a language model. MEMPHIS currently supports English, German and Italian, but the language identifier can easily be extended for more languages. Using a 300 character string from the middle of each document as input, the language is identified correctly for 99,5 % of 25000 test corpus documents from the MEMPHIS domains.

3. Document Classification

Within the MEMPHIS application, users can subscribe to certain topics. Statistical classifiers are used to assign one or more topics to newly acquired documents. The classification approach in MEMPHIS employs character-level n-grams with the naive-Bayes classifiers (Peng et al., 2003). Character n-grams are created by extracting all character sequences of length n from a text, treating all non-whitespace and whitespace characters equally, which means that word borders, punctuation, etc. may appear within an n-gram. We have experimented with n-grams of length 2 to 5.

Character-level n-grams have several advantages over term based n-grams often used in other statistical classification approaches like Rocchio, k-nearest neighbors, support vector machines and maximum entropy[1]. These approaches need some linguistic preprocessing that at least identifies the terms. Additionally, they suffer from the sparse data problem: Even with large training corpora,

[1] For an overview of statistical classification approaches see (Sebastiani, 2002).

there will be a significant number of terms in test data that are not contained in the training data. A common way to decrease sparse data is stemming. This requires an expensive linguistic preprocessing. Our approach needs no linguistic preprocessing at all and is completely language independent. This is especially important in MEMPHIS since we have input documents in different languages. Furthermore, documents may also contain spelling mistakes and/or no case distinction. The approach is robust enough to handle this. Finally, the amount of sparse data is much smaller than with term-based approaches. To handle the remaining sparse data, we adapted some standard smoothing techniques (Chen and Goodman, 1998).

Before classification takes place, language models must be trained. For each topic a corpus of relevant documents in each language is required from which a statistical language model is created. During runtime, a document is classified by calculating and comparing its probabilities for each model. We tested our approach on some of the corpora collected in MEMPHIS and achieved a classification accuracy above 90 %, depending on the classification parameters like training corpus size, n-gram length and smoothing technique. For a more extensive presentation of our classification approach with evaluation see (Steffen, 2004).

4. Shallow linguistic analysis

The basis for all subsequent processing steps is a shallow analysis of the source document which enriches it with linguistic annotation. For performing this task we utilize SProUT, a novel multilingual text processing platform (cf. (Drozdzynski et al., 2004)), equipped with a set of reusable online processing components for basic operations including tokenization, sentence identification, morphological analysis (including online compounding for German), gazetteer lookup, and reference matching. These resources can be flexibly combined into a pipeline that produces several streams of linguistically annotated structures which constitute an input for the shallow grammar interpreter, applied at the next stage for the identification of small-scale structures. Since the main motivation for developing SProUT centers around finding a trade-off between efficiency and expressiveness, the grammar formalism is a blend of very efficient finite-state techniques and unification-based type formalisms which are known to guarantee transparency and expressiveness.

A grammar consists of pattern/action rules, where the left hand side is a regular expression over typed feature structures (TFS) with functional operators and coreferences, representing the recognition pattern, and the right hand side is a TFS specification of the output structure. The usage of functional operators is twofold. First, they are used for forming the output of a rule (e.g., concatenation of strings) and secondly , they can act as predicates that produce boolean values, which can be used for introducing complex constraints. Coreferences express structural identity, create dynamic value assignments and serve as means of constraint propagation and information transport. Figure 1 shows a piece of grammar for recognition of location-PPs exemplifying the grammar formalism.

The first TFS matches a preposition. Then, one or zero determiners are matched. They are followed by zero or more adjectives. Finally, a location name (gazetteer) is consumed. The variables *#c*, *#n*, *#g* establish coreferences expressing the agreement in case, number, and gender for all but the last item (except for the preposition which solely agrees in case with the other items). The right hand side of the rule triggers the creation of a TFS of type *phrase*, where the surface form of the matched preposition and location are transported into the corresponding slots via the variables *#prep* and *#location*.

```
loc-pp :>
  morph & [POS Prep, SURFACE #prep,
           INFL [CASE #c]]
  morph & [POS Determiner,
           INFL [CASE #c,
                 NUMBER #n,
                 GENDER #g]] ?
  morph & [POS Adjective,
           INFL [CASE #c,
                 NUMBER #n,
                 GENDER #g]] *
  gazetteer & [TYPE general-location,
               SURFACE #location]
  -> phrase & [CAT location-pp,
               PREP #prep
               LOCATION #location].
```

Figure 1: A SProUT Rule

Within this highly declarative paradigm, we have developed grammars for recognition of persons, locations, organizations, temporal expressions and quantities, for the targeted languages. Interestingly, the grammars include explicit descriptions of how variant names of the recognized named-entities are built (e.g., a variant of a full person name might be last name). This information is utilized in order to discover mentions of previously recognized entities which turned to significantly boost the overall coverage. Named-entity identification provides additional level of analytic annotation, which is exploited by higher-level modules. In such a way, translation is improved considerably when the system can identify names as items not to be translated.

5. Summarization

For summarization a robust multi-lingual system was developed following a sentence extraction approach in the tradition of (Edmundson, 1969). The summaries created by the system have an indicative abstract quality which means that they are not intended to replace the original document, but rather indicate if the original document is of interest for a reader. The sentence extraction is based on a combination of several heuristics that are applied to the sentences of the document assigning them a relevance score. The result is the so-called *summary analysis*. From this analysis, summaries of different sizes can be created. Sensitivity to user provided query terms, so-called query-adaptivity, allows to generate different personalized summaries from the same text. This feature is used in MEMPHIS to provide personalized summaries by using the service and user provided subtopics from the user profiles as additional keywords to focus the summary on. As a side effect of the analysis, relevant keywords can be identified. In the following section, we give a short overview of the heuristics used. A more complete description can be found in (Kasper and Steffen, 2002).

5.1. The Term Weight Heuristic

Term weighting is based on a standard *tf.idf* approach (Salton and Yang, 1973), more precisely, the *atc* variant (Paijmans, 1997). The term frequency $tf(t)$ is the number of occurrences of term t in the input document. The inverse document frequency $idf(t)$ reflects the distribution of term t over a document corpus. The more documents in the corpus contain t, the lower is $idf(t)$. The idf of a term is retrieved from an idf database that must be built in advance, based on a training corpus. The tf.idf weight of a term is then defined as the normalized product of $tf(t)$ and $idf(t)$.

5.2. The Positional Heuristic

The positional heuristic exploits the different levels of headings and the paragraph structure of the document. The idea is that headings and the first sentence of a paragraph are more relevant for a summary than other sentences as has been shown in many summarization studies. Therefore, these sentences are given a higher relevance score by this heuristic.

5.3. The Layout Heuristic

This heuristic uses text mark-up like style, size and color as indicators of relevance. The idea is to exploit the fact that authors often change font properties to highlight or mark important phrases and text parts, and so this should be relevant for the summary, too. On the other hand, some text might be marked as unimportant, e.g. by using a font size smaller than the default.

6. Information Extraction

Information Extraction is the central functionality of the MEMPHIS system as it extracts from the documents retrieved that essential pieces of information the user is interested in. The extraction tasks are defined through templates to be filled. Documents consisting of free text require a semantic based analysis of the text based on the domain ontologies. Corpus based terminology extraction is used to identify relevant terms related to specific pieces of information and concepts. These terms are used as triggers in SProUT extraction grammars. For the financial news domain, e.g. names of companies, their stock prices, price offers and expected *return on equity (ROE)* of the companies involved in mergers have to be extracted.

For the MediAlert domain, a second approach was developed for semi-structured documents for which extraction rules based on document structure descriptions are used. For instance, it is difficult to identify titles of books or music, or that a certain name represents its author based on purely linguistic information. Site specific layout characteristics and mark-up provide valuable clues to that. This approach yielded very reliable results though it is not very robust and the document descriptions have to be updated when document structure changes. This happens frequently at the sites from which the media information is drawn. To support adaptation to new structures, a tool was created to learn such document descriptions from annotated examples. This tool generalizes on path expressions in the XML documents and generates test patterns. Another tool is used for detection of crucial structural changes at a site by regular cross-checking extraction results from known documents in order to alert the service provider for changes.

Besides the information types for representation to the service customers, additional meta-information is extracted from the documents to support the knowledge management system. For instance, in addition to the named entities recognized by the static SProUT grammars and gazetteers additional named entities with automatically generated variations for coreference matching can be acquired dynamically from the documents, e.g. through structural descriptions.

7. Machine Translation

Translation is used for translating summaries to the users' requested target language. The Logos MT system is used for this purpose. Though providing large coverage by itself, it requires adaptation to the service domains especially for terminology. Several thousand terms and rules were added for the Memphis domains.

7.1. Named Entity Recognition for MT

As one special difficulty in using the Logos MT system turned out that it is not capable to handle names except as unknown words which are not translated or by entering them into the dictionary which is not a general solution. Dangerous are the cases where the system identifies a name as a word it can translate, e.g. the German person name *Heidenreich* might be translated as *heathen empire*. Therefore, SProUT grammars for named entity recognition are used to preprocess the input text for MT to mark up terms not to be translated. As not all names should remain untranslated, e.g. many geographic names or names of international organizations need translation, a cross-check with the translation dictionaries is made to identify translatable terms. This shallow marking of terms provides a good intermediate solution for protecting terms. But a deeper integration with the MT system which would also allow to specify further linguistic information would be desirable, e.g. that the name is a certain type of NP.

8. Knowledge Management

8.1. Instance Matching

For the MediAlert application, the MEMPHIS system extracts information on book offers from many different providers. In order to integrate these different offers for the user into one comprehensive overview, a knowledge management component has been developed on the basis of the Protégé ontology development tool (cf. (Knublauch, 2003)). The component allows for the representation of different offers as a corresponding number of class instances, which will be merged into one instance if they overlap in title, authors and edition. This is checked through a small set of regular expressions that normalize case and spelling variations.

8.2. Knowledge Base Querying

The MEMPHIS system allows a customer to receive regular updates on recent book offers, optionally also according to a particular preference on topics (Tourism, Sports, etc.). In order to produce these from the knowledge base, a basic query application has been implemented. Querying extracts appropriate instances from the knowledge base according to constraints that correspond to user preferences (i.e. time period and topic). Additionally, the user may choose in which language he or she would prefer to receive information. At querying time, this preference is activated to produce a summarization of the book offer in the chosen language.

8.3. Knowledge Base Integration with External Ontologies

A further aspect of the knowledge management component is the possibility for integration of external ontologies into the MEMPHIS knowledge base. In this way, background knowledge on relevant aspects of the offers (e.g. tourist information for books on cities, countries) can be integrated into the book overview. By annotating the book summarizations with the external ontologies, it is possible to connect books according to related topics. The annotation introduces book offers as instances of ontology classes that are interconnected through relations defined in the ontology. If a customer requests information on a particular book offer, the system will now be able to provide many other related book offers.

8.4. Cross-lingual Retrieval

MEMPHIS allows user not only to choose from a predefined list of topics but also to add additionally terms as free text to further restrict the field of interest. If a user is interested in books about health he could also provide a term like “children” to indicate special interest in child health. Since these *user defined topics* are freely chosen one cannot use predefined classifiers to classify books for these topics. Also, as these terms are provided in the users’ native language, cross-lingual search capabilities are needed to identify matching book offers from other languages. At present, WordNet (c.f. (Miller, 1990)) and lookup of term translations in the Machine Translation dictionaries is used to provide sets of terms to search a document index across languages for book offers related to these user defined topics. This allows to find e.g. German books about “Kinderkrankheiten” within the health topic from “children” as user defined subtopic.

9. Conclusion

In the preceding sections, we presented a survey of core components of the MEMPHIS platform, focusing on functionality. But it should be obvious that each of these components has to be accompanied by a corresponding set of tools to create and adapt the application specific resources, whether linguistic resources, statistical models or domain specific knowledge. These tools, which make up the *MEMPHIS toolkit*, need to be easy to use and capable of adapting the MEMPHIS components to a large range of possible information services.

10. Acknowledgments

The work reported here is part of the MEMPHIS project funded by the European Commission under contract IST-2000-25045 in the *Information Society Technologies* (IST) program. Also, we would like to thank our research assistants here who provide invaluable support, especially Daniel Olejnik, Matthias Großklos and Eva Cissé.

11. References

Chen, Stanley F. and Joshua Goodman, 1998. An empirical study of smoothing techniques for language modeling. Technical Report 10-98, Harvard University.

Drozdzynski, W., H.-U. Krieger, J. Piskorski, U. Schfer, and F. Xu, 2004. Shallow processing with unification and typed feature structures: Foundations and applications. *German Journal of Artificial Intelligence (KI-Zeitschrift*, (1).

Dunning, Ted, 1994. Statistical identification of language. Technical Report CRL MCCS-94-273, New Mexico State University.

Edmundson, H.P., 1969. New methods in automatic extracting. *Journal of the ACM*, 16(2):264–285.

Kasper, Walter and Jörg Steffen, 2002. Multilingual flexible and robust summarization. In Stefan Busemann (ed.), *Konvens 2002. 6. Konferenz zur Verarbeitung natürlicher Sp rache. Proceedings*, number DFKI-D-02-01 in DFKI Document. DFKI GmbH, Stuhlsatzenhausweg 3, D-66123 Saarbrücken: DFKI GmbH.

Knublauch, H., 2003. An ai tool for the real world: Knowledge modeling with protege. *Javaworld, June 20, 2003.*

Miller, George A. (ed.), 1990. *WordNet: An on-line lexical database*, volume 3. International Journal of Lexicography 3, No. 4, 235-312.

Paijmans, H., 1997. Gravity wells of meaning: Detecting information-rich passages in scientific texts. *Journal of Documentation*, 53:520–536.

Peng, Fuchun, Dale Schuurmans, and Shaojun Wang, 2003. Language and task independent text categorization with simple language models. In Marti Hearst and Mari Ostendorf (eds.), *HLT-NAACL 2003: Main Proceedings*. Edmonton, Alberta, Canada: Association for Computational Linguistics.

Salton, G. and C. Yang, 1973. On the specification of term values in automatic indexing. *Journal of Documentation*, 29:351–372.

Sebastiani, Fabrizio, 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Steffen, Jörg, 2004. N-gram language modeling for robust multi-lingual document classification. In *The 4th International Conference on Language Resources and Evaluation (LREC2004)*. Paris, France: ELRA - European Language Resources Association. To appear.