[*First International Conference on Language Resources & Evaluation*, Granada, Spain, 28-30 May 1998]

Multilingual Information Processing: the AVENTINUS Project

Thomas Schneider
ts consulting

Abstract

The AVENTINUS project is supported by the EU Commission under the Telematics Program. Its aim is to provide software components and linguistic resources to drug enforcement authorities to improve multilingual communication and information processing. Components of the system are tools for translation such as machine translation, translation memory and terminology databases, information extraction modules identifying relevant object types (persons, drugs, places etc) based on a domain model, search engines for accessing both textual and structured data and large multilingual lexical resources, for both general vocabulary and drug and law enforcement terminology.
If the practical application of the system proves successful, extensions to other fields of organized crime are foreseen.

AVENTINUS is a multilingual information system for drug enforcement, which is partially funded by the European Commission under the Language Engineering program (LE1-2238).
For persons outside the field to better understand the structure of the project and the technological approach, some background information may be necessary.
The major reason behind the decision to fund the project was the seriousness of the situation in the areas of organized crime, drug trafficking, money laundering and terrorism. It has become painfully clear that drug dealing poses one of the greatest threats to the European nations. The damage attributable to drug dealing in Germany alone is estimated to be in the vicinity of 5 billion ECU per year, with exponential growth. If we were to extrapolate and project this local damage to the World community, the figures would be staggering.
It is self-evident that drug enforcement must be an internationally coordinated action. A single national effort is likely to fail since criminal networks operate on an international scale (and the problem has increased after the break-up of the Soviet Union and the opening of the eastern borders of the EU).

It is not that the European police agencies and government institutions have not been cooperating. But to this day there are some major obstacles to efficient information exchange and fast response time. First of all, there may be relevant information available, but the officer may not know where to look. This information may be spread over several different information sources, e.g. chemical descriptions of substances, photographs, police reports, person files or video sequences. And this information resides on different computers in different countries - and in different languages. Not everybody speaks the language in which the required information is available. It is possible today to transmit data packages in fractions of a second, but if the content is incomprehensible, the speed is irrelevant.

Project Description

In contrast to most research projects, AVENTINUS is strictly user-driven. The members of the user group are responsible for the definition of requirements. Only they have the background to decide which technical solutions can contribute to their practical work. Not only do all technical specifications have to be approved by the user group, but any existing or newly developed tool must fit into their existing infrastructure. No agency can afford to scrap in-house IT systems which have been developed and loaded with data over many years in favor of radically different tools and approaches. In some cases, using newer technologies might provide efficiency gains, but their introduction would severely disrupt internal workflow and established procedures.
Another aspect not to be forgotten is that most of the data handled by drug enforcement authorities are extremely sensitive and confidential and may not be made available to outsiders. For researchers and developers this means that most of the hands-on material available is toy data - which of course hampers precise focussing of development and creates a dilemma in planning the practical applications.
In general, there are two scenarios which need to be supported: The first one involves the acquisition and interpretation of data. In each of the government agencies, every day many items of information are received: plain text, formatted messages, telexes, faxes or video sequences, and all this from a wide range of different sources and in different languages. The first step (which is outside of the AVENTINUS project) is to convert paper documents into machine-readable text. Next is the attempt to identify the language involved, not always a trivial task if your expertise is not in Asian or Eastern European languages. Then the text has to be translated by one of the translators available for the language, and the content has to be described and stored in a database for future use. This is a lengthy process which requires not only a lot of time but also considerable resources.

Within AVENTINUS. the process will be automated as far as possible. A computer program will automatically identify the language involved. If a machine translation system is available for the language, the document will be channeled to that system. As an alternative, the document will pass by a translation memory, or if that proves ineffective, a terminology database. In the latter case, terms detected in the database are inserted into the original text. This cannot replace a translation but it can give the analyst at least a hint as to the content. This might be especially helpful if the text is written in a language for which no machine translation system is available, as for e.g. Arabic. Based on this assessment of relevance, it is easier to decide whether a given text warrants translation or is less urgent.

The second scenario involves the retrieval of relevant information for the case at hand. Agents of drug enforcement units need quick access to data, be it information on a person's movements, properties of a certain drug or current legislation on narcotic substances. Usually this information is available somewhere, but retrieval of the relevant facts is time-consuming. Conventional techniques of querying full-text databases produce more "noise" than accurate hits, and in the past the officer had to be proficient in the language of the source to utilize it. In a large database of perhaps millions of text passages, a search based on key words is likely to produce more answers than can be looked at in detail. On the other hand, the same concept might be expressed in a different surface form, with the relevant concept described perhaps in a subordinate clause or expressed by a synonym, and probably not be found at all. A user would actually have to know the possible expressions in advance to have a chance of finding the texts. Within AVENTINUS, a linguistic analysis of the text coupled with an automatic generation of conceptual links between content words based on the statistical analysis of the contextual occurrence of the words produces a much higher percentage of correct and focused answers. A module supporting fuzzy search will also permit finding misspelt words or names which had been transliterated in different ways, e.g. names taken from Cyrillic are usually represented differently in German than in English.

To query a foreign language database, in the past a user had to express his questions in that foreign language, again trying to anticipate potentially different expressions of the concepts, a time-consuming and error-prone procedure.

AVENTINUS will, by contrast, permit an officer to formulate a query in his own native language. Invisible to the user, the question will be translated into the appropriate form and the appropriate language for the various databases. Some data such as chemical tables are stored in structured form; others are hidden in running text. By using sentence level syntactic and lexical analysis, building dependency relations between concepts, AVENTINUS will either retrieve the relevant documents while avoiding to overload the user with digital garbage, or extract the relevant facts from the texts stored in one of the databases. Fact extraction is a highly complex problem as it requires the development of intricate but still practically manageable domain models, hyperlinks between concepts and access to databases with e.g. named entities. Machine translation, translation memory and term substitution systems will see to it that the user receives the requested information in his own native language. In other words, AVENTINUS will combine information from structured as well as unstructured databases into an understandable package even if the sources reside in different countries and are formulated in different languages. This approach should speed up the search of relevant information and greatly improve accuracy and coverage.

Application Domain

The data handled within the AVENTINUS project are of several different types, the most important ones being textual data. Different types of text have to be considered:

- legal texts describing rules and procedures in drug enforcement but covering also court decisions, legal frameworks for data exchange or permissible functions of organizations like EUROPOL. These legal texts are usually available in several languages and can be treated with Translation Memories.
- texts from open sources (such as newswire texts). Such texts have to be assessed as to their relevance for drug enforcement and must be processed with information extraction and indexing techniques
- Internal communication texts such as police reports. Usually these texts are structured but they may also contain a fair amount of free text. However, as they are generally standardized according to content and style, they can be translated automatically with a fairly high degree of precision.

Workflow and Processing

For the "incoming data" scenario sketched above, it must be decided whether a given input document is relevant or not. So the task is threefold:

- identify drug related texts in any given language
- make foreign language texts at least understandable to the degree that users can guess their relevance
- apply information extraction and indexing techniques to facilitate further processing

To this end, the project will provide translation support tools and tools for information understanding.

For the "fact retrieval" scenario, the user's search task must be supported. While the most widely used search objects are proper names, places and dates, other information e.g. in open sources must be targeted as well.

The project will provide tools like name search, including transliterations and references to similar names, and text search in both structured and free text databases. Translation support tools are responsible for the translation of both the search requests and the search results. Multimedia objects are considered if they are indexed

Components

The further development, adaptation and integration of relevant components are the main focus of the AVENTINUS development tasks.

Translation Support

Within the project, three types of translation support tools are used. All of them are available as stand-alone tools accessing common lexical resources and can be called from standard editors such as WinWord.

The **Term Substitution** component looks up foreign words in a terminology database and replaces them in the text with the respective native language terms. In many cases, a foreign language text can be made marginally understandable in this way.

Term substitution is the most robust translation technology. It uses basic tools like taggers and lemmatizers for the foreign language and looks up single words as well as multi-term words in the lexicon. If a word is found, it is either written into a list of known terms or is inserted into the input text in a different color for highlighting.

The **Translation Memory** is useful for the treatment of structured and standardized texts such as police reports. The component used in AVENTINUS is PC-TM developed by ILSP. It has sophisticated alignment capabilities to build new memories; it is able to run database search both interactively and in batch mode; it supports different levels of both perfect and fuzzy matches. Moreover, it can be trained for variable matches if combined with some information extraction techniques so that named entities can be treated as variables in the context of otherwise perfect matches. The translation memory is available both as a client server and a stand-alone version under Windows.

A **Machine Translation** component is integrated for the treatment of texts which are not as standardized as police reports. Since the aim is to gather information rather than produce high-quality output for publication purposes, a sophisticated MT component is likely to produce acceptable translations. However, as MT is not available for all languages that need to be considered, coverage can only be partial.

The system integrated in AVENTINUS is the T1 system of GMS marketed by Langenscheidt, based on the former Siemens METAL translation system. It is now available on PC in standard object oriented technology and can be used from a standard Windows editor.

As all components use the identical lexical resources, great care had to be taken in the design of the common database so that no overhead is created for users in exchanging terminology between the central lexicon and the more specialized lexicons of the MT component.

Information Processing Support

For information extraction, there is special interest in recognizing the following named entities: narcotics, persons, criminal and non-criminal organizations, transportation means and routes, communications means, places and dates. Targeted texts for this type of search are newswire messages and police reports. A second level of information extraction is template generation, e.g. on persons and organizations. These templates are presented in a standard form with which users are familiar already, for example the forms used by INTERPOL. In the framework of the project, the Vanilla Information Extraction technology provided by the University of Sheffield is adapted to areas which are specific for AVENTINUS and will be extended into a multilingual environment. French, German and Spanish components are under development.

For indexing support, close interaction is required with the retrieval model. Here, the legacy from the past is that the different users have implemented different systems over the years, and unfortunately, the largest common denominator for all users are Boolean techniques. So there will be a two-level indexing component:

- local indexing, meaning that all possible information from a document is extracted without reference to the system content; especially keyword information and some frequency information is provided
- contextual indexing, meaning that some contextual features (term weights, dependency relations between terms etc) are exploited. Within the restricted domain, thesaurus information can be used profitably as well.

Indexing is based on the work of (Ruge, 1992) and (Ruge; Schwarz; Warner, 1991) and is being extended to the languages covered in AVENTINUS.

As stated above, the **Search Support** covers several requirements:

- Users can forward their search request in natural language as well as in some structured form (e.g. by filling in templates)
- Users can forward their search request in their native language instead of the foreign languages of the database to be searched
- Users have query expansion and navigation possibilities
- As it is not a priori obvious whether a search request will lead to a search in a structured database or a textual database, both options are supported.

The Name Search component offers the possibility to look for similar names. There are several techniques to identify similar names, mainly the somewhat language-independent N-gram similarity measuring combined with some simple phonetic-based algorithms. These results can still be improved by techniques for language-specific similarity checking which refer to e.g. transcriptions of Arabic or Russian names and potential object identity. Prototypes for name recognition have been developed by GMS.

Search support in texts will be at two levels, first by launching some recall-related search by expanding the search requests and then by adding some precision-related operations on the output for ranking. Focus will be on user interaction as conceptual mismatches between users and the systems have been proven to be responsible for a good deal of low retrieval results.

- Users have the possibility to state a structured query, e.g. by filling in forms, even if textual material is searched
- Users forward textual queries in their own native language. A multilingual terminology database is used as a filter for the search request; it can be further tuned towards specific domain restrictions to avoid any noise in the search
- Users are provided with a thesaurus containing links between terms (produced both manually and automatically) to permit the selection of the optimal search terms

Documents retrieved will be ranked; then the translation tools described above are applied to make the documents understandable in the user's native language.

The **Structured Database Search** will transform the natural language query into an SQL statement. Due to the type of data to be searched and the type of user (who is no linguistic expert) a simple pattern-based solution will be used. Focus will be put on easy user interaction by providing intelligent feedback on how the system understood a structured search request. Input will be both by query-by-example and by natural language search requests. The effectiveness of the one method versus the other needs to be determined in user tests.

Resources

Three types of linguistic resources are being built up. Foremost, there is the multilingual lexical database of drug terminology and related areas. Each record contains the term, its part of speech, a definition of controlled terms using a semantic hierarchy and the target language equivalents. At present, it comprises several thousand entries. In addition, there will be special purpose lexicons needed by the information extraction or linguistic processing components, e.g. place names, titles, abbreviations etc, as well as common vocabulary in different languages. The specific resources needed for the project are being collected by the user partners, while significant lexical and terminological resources are already available by the development partners.

The **Thesaurus** will store all information available for a given term and links to other terms. It will provide the classical information retrieval links such as Broader term/Narrower term, but also additional linguistically motivated links, like morphological similarity, semantic similarity, head-modifier links etc. The thesaurus, which is implemented as part of the general lexicon, will also have multilingual links. Modules like the information extraction components or the structured query generation component need access to a domain model. It must be made multilingual and be connected to the lexicon information. Words have to be linked to concept nodes (which is analogous to problems in MT transfer), and the lexicon entries need to be classified as to the concept links they can undergo.

Architecture

The system's architecture must follow two basic principles. It must be based on components which can be integrated in a very flexible way into the existing system environment of the users. This requirement implies a maximum of modularity in interfaces and in its use of internal resources. Internal components must be available at different phases of the workflow, as no predefined processing sequence can be given.

As a consequence, the architecture consists of a framework which is defined at three levels: text level, user interface level and access to resources. All tools and components use the same text mark-up format. This format is SGML based and defines a common structure for all textual properties which are linguistically relevant. For outside formats such as RTF or ASCII, converters will be provided.

The common linguistic resources are accessed via a common API. The same applies to other common functions and procedures such as taggers or lemmatizers. As potential users may decide not to use all components, they have to be clearly separated and be able to operate as stand-alone units.

Another aspect to be considered is the ease of administration. Resources must be in a coherent state for all linguistic components. One update of resources must be propagated to all system components; multiples resource administration must be avoided. At the same time, care has to be taken that users are not subjected to tasks in which they have no expertise, such as adding detailed linguistic descriptions to new entries.

The first version of the overall system will soon be beta-tested at the EUROPOL DRUGS UNIT. If it proves successful in its practical application, an extension to other areas is planned. Porting the approaches to other fields such as economic intelligence will require the adaptation of the tools and components and the collection of resources ranging from domain models and multi-functional lexicons to specialized terminology.

References

Boutsis,M., Piperidis, St.(1996). Automatic Extraction of Bilingual Lexical Equivalences from Parallel Corpora. *Proceedings ECAI 1996*

Cunningham,H., Gaizauskas,R., Wilks,Y.(1996): GATE - a General Architecture for Text Engineering. *Proceedings COLING, Copenhagen 1996*

Ruge,G.(1992):Experiments on Linguistically Based Term Associations. *Information Processing and Management 28(3) , pp.317-332*

Ruge.G., Schwarz,C., Warner,A. (1991): Effectiveness and Efficiency in Natural Language Processing for Large Amounts of Text. *JASIS 42(6), pp 450-456*

Thurmair, G.(1990): METAL - Computer Integrated Translation. *Proceedings SALT Workshop,* Manchester 1990

Thurmair, G.(1997): Information Extraction for Intelligence Systems. *Proceedings of the Conference on Natural Language Processing: Extracting Information for Business Needs. unicom:London 1997*