University of Hagen at CLEF 2005: Towards a Better Baseline for NLP Methods in Domain-Specific Information Retrieval

Johannes Leveling
Intelligent Information and Communication Systems (IICS)
University of Hagen (FernUniversität in Hagen)
58084 Hagen, Germany
Johannes.Leveling@fernuni-hagen.de

Abstract

The third participation of the University of Hagen at the German Indexing and Retrieval Test (GIRT) task of the Cross Language Evaluation Campaign (CLEF 2005) aims at providing a better baseline for experiments with natural language processing (NLP) methods in domain-specific information retrieval (IR).

Our monolingual experiments with the German document collection are based on a setup combining several methods to achieve a better performance. The setup includes an entry vocabulary module (EVM), query expansion with semantically related concepts, and a blind feedback technique. The monolingual experiments focus on comparing two techniques for constructing database queries: creating a *'bag of words'* and creating a semantic network by means of a syntactico-semantic parser for a deep linguistic analysis of the query. The best performance in the official experiments was achieved by a setup using staged logistic regression, a query expansion with semantically related concepts, an entry vocabulary module, a deep linguistic analysis of the query, and blind feedback (0.2875 mean average precision (MAP)). Additional experiments showed a performance improvement when changing to the basic Okapi BM25 search (0.3878 MAP).

For the bilingual experiments, the English topics are translated into German queries with several machine translation services available online (Systran, Free translation, WorldLingo, and Promt). Each set of translated topics is processed separately with the same techniques as in the monolingual experiments. The best performance was achieved with a query translation by Promt with a simple keyword extraction from the translation (0.2399 MAP with a staged logistic regression approach vs. 0.2807 MAP with Okapi BM25).

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods, Linguistic processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Query formulation, Search process*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*; I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods—*Semantic networks*

General Terms

Experimentation, Performance, Measurement

Keywords

Natural language processing for information retrieval, Deep linguistic analysis, Keyword extraction

1 Introduction

This paper presents the work for the third participation of the University of Hagen in the domain-specific GIRT task in the evaluation campaign of the Cross Language Evaluation Forum (CLEF). Natural language processing (NLP) as described in the following subsections is part of query processing for the NLI-Z39.50[1] (Leveling and Helbig, 2002), a natural language interface for databases supporting the Internet protocol Z39.50 (ISO, 1998). A major part of our experimental infrastructure was developed for and is applied in this NLP system.

In CLEF 2003, retrieval strategies based on generating query variants for a single natural language (NL) topic were compared (Leveling, 2004). The best experiment of the University of Hagen in the domain-specific task in 2003 with respect to mean average precision (MAP) used multiple variants of a query, a query expansion with semantically related concepts, and a database index containing word forms (0.2064 MAP for a run using both topic title and description).

In CLEF 2004, the focus of our experiments was on investigating differences in indexing methods, such as indexing unprocessed word forms, indexing concepts, and indexing semantic networks (Leveling and Hartrumpf, 2005). The best experiment of the University of Hagen in the domain-specific task in 2004 with respect to mean average precision used a single query, query expansion with semantically related concepts, and a database index containing word forms (0.2482 MAP for a monolingual German run using both topic title and description).

For the monolingual experiments in CLEF 2005 two main objectives are pursued:

1. To establish a better baseline for a comparison between NLP methods and traditional approaches in IR and to achieve a better overall performance. To complete this goal, some methods from the experimental setup of the UC Berkeley in previous years were adapted.

2. To compare two techniques for creating database queries from the natural language topics: a) extracting keywords (*'bag of words'*) and b) applying a deep linguistic analysis by means of a syntactico-semantic parser before creating a database query.

For the bilingual experiments with the GIRT collection, the English query topics are translated into German by means of several free machine translation services that are available over the Internet. The translations obtained from each machine translation service are processed in separate experiments, employing the query processing techniques for monolingual experiments described above.

1.1 Towards a Better Baseline

To state and investigate a hypothesis such as *"The use of semantic networks (or NLP in general) – including methods such as expanding queries with semantically related concepts – improves performance in IR."* an acceptable experimental baseline has to be established.

The experimental setup used by the University of Hagen in previous participations at CLEF did not provide a performance higher than other state-of-the-art systems. It served mainly as a basis for comparing different strategies on the same system (intra-system comparison) and did not typically aim at an overall superior performance (inter-system comparison).

To obtain a better baseline for our experiments, several methods of the experiments of the UC Berkeley for CLEF were added to our experimental setup to improve retrieval performance, including a so-called entry vocabulary module. In total, our setup now includes the following options:

[1]The NLI-Z39.50 is being developed as part of the project "Natürlichsprachliches Interface für die internationale Standardschnittstelle Z39.50" and funded by the DFG (Deutsche Forschungsgemeinschaft) within the support program for libraries "Modernisierung und Rationalisierung in wissenschaftlichen Bibliotheken".

- LA: extracting keywords (*'bag of words'*) or applying a deep linguistic analysis, i.e. employing NLP methods to produce a semantic network representation (described in (Leveling, 2004; Leveling and Hartrumpf, 2005)).
- EVM: employing an entry vocabulary module. An entry vocabulary module provides a mapping of words from a possibly uncontrolled vocabulary to a controlled vocabulary, based on likelihoods of co-occurrence (Gey and Chen, 1997; Gey et al., 2001b,a). As suggested in (Petras, 2005), the top five ranked terms from the EVM are added to the database query, down weighted by half of the default weight for our experimental runs.
- BF: using blind feedback. The top-N controlled terms from the top-M ranked documents are extracted for a query reformulation (Petras et al., 2004; Petras, 2005). As suggested in (Petras et al., 2004), thirty terms from the top twenty documents are extracted and down weighted by half of default weight in our experiments ($N = 30, M = 20$).
- QEX: adding semantically related concepts to the query. The query expansion stage is based on adding semantically related concepts (word senses) to a query, including synonyms, hyponyms, meronyms. This approach was described in (Leveling, 2004).

In addition, our database setup consists of:

- The Cheshire II database (Larson and McDonough, 1997; Larson et al., 1996), which supports Boolean searches, probabilistic weighting with staged logistic regression (Cooper et al., 1992, 1994), which is used in the UC Berkeley setup as a default, and basic Okapi BM25 (Robertson et al., 1994) (opposed to the Zebra database used in our earlier experiments; Zebra supports Boolean searches and the standard *tf-idf* weighting as a default).
- A document representation which results from applying
  - the WOCADI parser (WOrd ClAss based DIsambiguating parser) to obtain results for the morpho-lexical analysis for the title and abstract (opposed to indexing unprocessed word forms as in our experiments for CLEF 2003),
  - a stopword list of a few hundred word forms (opposed to indexing all word forms, concepts, or semantic networks as in our experiments for CLEF 2004),
  - a German stemmer, which was originally implemented in Snowball[2] (opposed to indexing word forms or lemmata), and
  - a linguistically motivated (lexicon-based) decomposition of German noun compounds.

1.2 Techniques for Query Processing

Two techniques for query processing are compared. The first technique corresponds to extracting keywords from the topic title and topic description to create a database query. The query string is tokenized and word forms are extracted. Some normalization steps such as stopword removal and stemming are employed to produce a database query in the Database Independent Query Representation (DIQR, see (Leveling, 2004)). This roughly corresponds to the traditional approach of *'processing'* natural language queries in information retrieval (*'bag of words'*).

The second query processing technique employs NLP methods to create a database query. A syntactico-semantic parser, WOCADI (Hartrumpf, 2003), produces a semantic network representation of the query according to the MultiNet paradigm (Helbig, 2001, 2005) which is then transformed into a DIQR by means of a rule-based transformation engine consisting of a Rete compiler and a Rete interpreter (the implementation is described in more detail in (Leveling and Helbig, 2002)).

For both techniques, the DIQR is mapped to a query in a formal language the database management software supports (such as a query for the Z39.50 protocol) which is then submitted to the target database.

[2] http://snowball.tartarus.org/

Table 1: Overview of parameter settings and results for official and additional monolingual GIRT experiments with the German document collection. The results displayed are the mean average precision (MAP), the number of relevant and retrieved documents (rel_ret), and the total number of retrieved documents for 25 topics. Results of the official runs are set in *italics*.

Run Identifier	Parameters				Staged logistic regression			Okapi BM25		
	QEX	EVM	LA	BF	MAP	rel_ret	# docs	MAP	rel_ret	# docs
FUHggnnn	no	no	no	no	0.2825	1715	23782	0.3522	1939	23782
FUHggnnl	no	no	yes	no	0.2972	1776	24724	0.3478	1914	24724
FUHggynn	yes	no	no	no	*0.2859*	*1849*	*25000*	0.3188	1972	25000
FUHggynl	yes	no	yes	no	0.2477	1818	25000	0.2808	1988	25000
FUHggnyn	no	yes	no	no	0.2540	1596	24058	0.3168	1867	24058
FUHggnyl	no	yes	yes	no	0.2728	1679	24724	0.3043	1864	24724
FUHggyyn	yes	yes	no	no	*0.3017*	*1882*	*25000*	0.3587	2093	25000
FUHggyyl	yes	yes	yes	no	0.2876	1898	25000	0.3543	2099	25000
FUHggyydbf	yes	yes	no	yes	*0.3031*	*1882*	*25000*	0.3701	2141	25000
FUHggyylbf	yes	yes	yes	yes	0.3017	1923	25000	**0.3878**	**2207**	**25000**

2 Monolingual GIRT Experiments (German – German)

2.1 Experimental Setup and Results

For the GIRT task in 2005, several experimental runs for the monolingual GIRT task were submitted. The experiments vary in the following parameter settings: using a query expansion with semantically related terms (QEX=yes/no), using an entry vocabulary module (EVM=yes/no), constructing a query from the semantic network after a linguistic analysis or using a traditional keyword extraction (LA=yes/no), and using blind feedback (BF=yes/no). Table 1 gives an overview of the monolingual experiments performed for the German document collection with their results.

2.2 Discussion of Results

The best monolingual GIRT experiment used a query expansion with semantically related concepts, an entry vocabulary module, and blind feedback. The performance of the best official experiment with respect to mean average precision (0.3031 MAP for run FUHggyydbf) is better in comparison to our experiments in CLEF 2003 (0.2064 MAP) and CLEF 2004 (0.2482 MAP). Additional experiments were conducted using the state-of-the-art weighting approach, Okapi BM25, which yield a significantly better performance (0.3878 MAP for run FUHggyylbf).

The effect of any single query processing method (corresponding to a single parameter) is still inconclusive, but the combination of all processing methods with a deep linguistic analysis of the query yields the best performance with respect to the number of relevant and retrieved documents and mean average precision (FUHggyylbf).

Several differences between our setup and the experimental setup of the UC Berkeley still remain and may account for a difference in performance.

- Morphological processing:
  In the setup of the UC Berkeley, the stemmer is applied to word forms. In our setup, stemming is applied after lemmatizing. The WOCADI parser returns morpho-lexical results for the sentences, including the lemmata (complex named entities and multi-word lexemes are not identified in the morpho-lexical analysis at the moment). The input data for the German stemmer consists of these lemmata from the morpho-lexical analysis of a sentence, i.e. the stemming process starts with normalized word forms. Plural forms of nouns, participles, past forms of verbs and verb with a separable prefix are already normalized.

- Database Index:
  In the setup of the UC Berkeley, only the constituents of German compounds are used as index terms. In our setup, compounds are indexed together with their constituents (because of a later query expansion with semantically related compounds). In addition, our database index contains stemmed word forms from the title and the abstract fields of the documents only; a database index containing subject terms as well was not created.

- Phrases:
  In our setup, a simple frequency-based phrase recognition for the EVM is employed which identifies adjective-noun sequences (English noun phrases often correspond to compounds in German).

- Compounds:
  In the UC Berkeley setup, decompounding of compounds is based on a statistical approach (Chen, 2002). In our setup, decompounding is lexicon-based and includes a solution of the so called Fugen-problem (omitting, inserting, or substituting letters in or between constituents for a German compound, as an omitted *'e'* in *'Schulsprecher'* – *'Schule'* + *'Sprecher'*), an additional *'s'* (*'Verfahrensfehler'* – *'Verfahren'* + *'Fehler'*), or an Umlaut in the plural form (*'Gänsefleisch'* – *'Gans'* + *'Fleisch'*).[3] Compounds which should not be split into their constituents are entered into the computational lexicon to block decompounding (e.g., *'Frauenzimmer'*/*'dame'* does not have regular semantic relations to its constituents and is represented as a concept).

- Entry Vocabulary Module:
  In our setup, the EVM uses a co-occurrence between title words and subject terms only and does not include words from the abstract.

These differences can account for a significantly lower performance of our experiments. For example, we conducted tests which showed a different performance for indexing lemmata and indexing stemmed word forms. Experiments on the index with stemmed word forms resulted in a significantly better performance because the stemmed forms conflate adjectives and nouns into the same form (e.g. *'wirtschaftlich'*/*'economical'* and *'Wirtschaft'*/*'economy'* are conflated into *'wirtschaft'*/*'economi'*). For a database index with lemmata, only inflectional endings are removed and e.g. *'wirtschaftlich'* and *'wirtschaft'* remain different index terms. While treating different concepts differently remains one of our objectives for a high precision information search, our background knowledge which semantically links the correct senses of *'Wirtschaft'* and *'wirtschaftlich'* is neither robust (i.e., it relies on correct word sense disambiguation) nor lexically complete, i.e. the linking between corresponding concepts (derivational links) in our computational lexicon is not complete, yet.

3 Bilingual GIRT Experiments (English – German)

In CLEF 2004, the bilingual experiments of the University of Hagen were based on a method which combines a concept translation using linguistic resources (such as GermaNet and EuroWordNet) with a word translation using translation lists to obtain a ranked list of translation alternatives. This method was employed to translate German concepts in the DIQR into English to create a database query. For CLEF 2005, English query topics are submitted to freely available machine translation services for a translation into German queries.

3.1 Experimental Setup and Results

Our second participation in the bilingual GIRT task (matching English topics against the German data) is based on various machine translation services for a translation. For the bilingual retrieval experiments

[3] A *'Fuge'* does not correspond to a morphological suffix and can not be treated by a suffix elimination process (stemming).

Table 2: Overview of parameter settings and results for bilingual GIRT experiments with English queries and the German document collection. The results displayed are the mean average precision (MAP) and the number of relevant and retrieved documents (rel_ret). For each run, 25 000 documents were retrieved. Results of the official runs are set in *italics*.

Run Identifier	Parameters					Staged logistic regression		Okapi BM25	
	Translation	QEX	EVM	LA	BF	MAP	rel_ret	MAP	rel_ret
FUHegfyyn	Free translation	yes	yes	no	no	*0.2392*	*1518*	0.2639	1762
FUHegsyyn	Systran	yes	yes	no	no	*0.2261*	*1584*	0.2639	1762
FUHegwyyn	WorldLingo	yes	yes	no	no	*0.2248*	*1580*	0.2639	1762
FUHegpyyn	Promt	yes	yes	no	no	*0.2399*	*1567*	**0.2807**	**1767**
FUHegpyyl	Promt	yes	yes	yes	no	0.2111	1518	0.2447	1733

(English – German) with the GIRT document collection, four machine translation services translate the English topics into German queries: Free translation[4], Systran[5], and WorldLingo[6], and Promt[7].

For the official runs, keywords are extracted from the translated queries: the query is tokenized, stopwords are removed, and a stemmer is applied. The remaining stemmed words are employed to create a simple DIQR, which is processed accordingly. For a fifth experiment, the topics are translated by the Promt translator, analyzed by WOCADI to produce a semantic network representation, and transformed into a DIQR with the Rete-based transformation. The DIQR queries are then processed as described in Section 2. Table 2 gives an overview of the bilingual experiments with English topics on the German document collection together with their results.

3.2 Discussion of Results

The WOCADI parser may serve as an indicator for the quality of the translated queries. WOCADI performs many morpho-syntactical and semantical tests while trying to produce a semantic network representation for a query topic. If some of these tests (such as agreement between subject and verb, selectional restrictions for the complement of an action (verb), etc.) fail or if the translation contains other errors, a partial semantic network is returned. For severe errors in the translations, no network is returned at all.

WOCADI reported no problems in parsing the original German topics. Parse results for the translated German queries showed that many of the 200 translations (25 topics, four different machine translations services, and translation of title and description) contained a syntactical or lexical error, so that WOCADI could not fully analyze these translations. For example, all services had problems translating the imperative forms of the verb *'find'*. (The exclamation mark to indicate the imperative is missing in all topics and the original queries were not modified by us.)

Due to these test results, we judge the quality of online translations so poor that the results from translating English topics into *German* can not be processed successfully with a deep linguistic analysis. Thus, the effect of trying to parse a grammatically incorrect translation (a partial semantic network or no semantic network is returned by WOCADI) is not present in a simple keyword extraction from a shallow analysis (the morpho-lexical stage in the WOCADI parser). All potential advantages of a deep linguistic processing are then unavailable (FUHegpyyn vs. FUHegpyyl). A performance increase was observed for all experiments using Okapi BM25.

[4] http://www.freetranslation.com/
[5] http://www.systransoft.com/
[6] http://www.worldlingo.com/
[7] http://www.e-promt.com/en/

4 Conclusion

In comparison with the results for the monolingual GIRT task in 2003 and 2004, performance with respect to the MAP for the best official experiment has improved considerably (0.2064 MAP in 2003, 0.2482 in 2004 vs. 0.3031 in 2005). Additional experiments involved changing the ranking scheme to Okapi BM25, which increased the number of relevant and retrieved documents and the mean average precision significantly (0.3878 MAP). As indicated by the better performance, the setup for the additional experiments provides a much better baseline for experiments with NLP methods in IR.

The method for constructing a database query using the transformation of the semantic network representation into a database query yields a higher performance than extracting keywords in combination with all other methods applied in these experiments. Results are still inconclusive in which cases NLP methods provide a better performance and even seem to depend on the ranking scheme employed.

The machine translation services tested did not produce high-quality translations. At the moment, using a robust keyword extraction yields better performance than a subsequent try to semantically analyze malformed *'translations'*.

The experimental setup of the University of Hagen and of the UC Berkeley still differ much, despite using methods which are comparable, similar or even the same with respect to some experimental parameters for the monolingual experiments. Further research is required and should properly identify the differences in these parameters and their influence on the overall retrieval performance.

References

Chen, Aitao (2002). Multilingual information retrieval using English and Chinese queries. In *Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001* (edited by Peters, C.; M. Braschler; J. Gonzalo; and M. Kluck), volume 2406 of *Lecture Notes in Computer Science (LNCS)*, pp. 44–58. Berlin: Springer.

Cooper, William S.; Fredric C. Gey; and Aitoa Chen (1994). Probabilistic retrieval in the TIPSTER collections: An application of staged logistic regression. In *Proceedings of the Second Text REtrieval Conference (TREC-2)* (edited by Harman, D.), pp. 57–66. National Institute of Standards and Technology (NIST), Special Publication 500-215.

Cooper, William S.; Fredric C. Gey; and Daniel P. Dabney (1992). Probabilistic retrieval based on staged logistic regression. In *SIGIR '92: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 198–210. New York, NY, USA: ACM Press.

Gey, Fredric C.; Michael Buckland; Aitao Chen; and Ray R. Larson (2001a). Entry vocabulary – a technology to enhance digital search. In *Proceedings of the First International Conference on Human Language Technology*. San Diego.

Gey, Fredric C. and Aitao Chen (1997). Phrase discovery for English and cross-language retrieval at TREC 6. In *Proceedings of the Sixth Text REtrieval Conference (TREC-6)* (edited by Voorhees, E. M. and D. Harman), pp. 637–647. National Institute of Standards and Technology (NIST), Special Publication 500-240.

Gey, Fredric C.; Hailing Jiang; Vivien Petras; and Aitao Chen (2001b). Cross-language retrieval for the CLEF collections – comparing multiple methods of retrieval. In *Cross-Language Information Retrieval and Evaluation: Workshop of Cross-Language Evaluation Forum, CLEF 2000* (edited by Peters, C.), volume 2069 of *Lecture Notes in Computer Science (LNCS)*, pp. 116–128. Berlin: Springer.

Hartrumpf, Sven (2003). *Hybrid Disambiguation in Natural Language Analysis*. Osnabrück, Germany: Der Andere Verlag. ISBN 3-89959-080-5.

Helbig, Hermann (2001). *Die semantische Struktur natürlicher Sprache: Wissensrepräsentation mit MultiNet*. Berlin: Springer.

Helbig, Hermann (2005). *Knowledge Representation and the Semantics of Natural Language*. Berlin: Springer.

ISO (editor) (1998). *ISO 23950:1998, Information and documentation – Information retrieval (Z39.50) – Application service definition and protocol specification*. Geneva: International Organization for Standardization.

Larson, Ray R. and Jerome McDonough (1997). Cheshire II at TREC 6: Interactive probabilistic retrieval. In *Proceedings of the Sixth Text REtrieval Conference (TREC-6)* (edited by Voorhees, E. M. and D. Harman), pp. 649–660. National Institute of Standards and Technology (NIST), Special Publication 500-240.

Larson, Ray R.; Jerome McDonough; Paul O'Leary; Lucy Kuntz; and Ralph Moon (1996). Cheshire II: designing a next-generation online catalog. *Journal of the American Society for Information Science*, 47(7):555–567.

Leveling, Johannes (2004). University of Hagen at CLEF 2003: Natural language access to the GIRT4 data. In *Comparative Evaluation of Multilingual Information Access Systems, 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Revised Selected Papers* (edited by Peters, Carol; Julio Gonzalo; Martin Braschler; and Michael Kluck), volume 3237 of *Lecture Notes in Computer Science (LNCS)*, pp. 412–424. Berlin: Springer.

Leveling, Johannes and Sven Hartrumpf (2005). University of Hagen at CLEF 2004: Indexing and translating concepts for the GIRT task. In *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign* (edited by Peters, C.; P. D. Clough; G. J. F. Jones; J. Gonzalo; M. Kluck; and B. Magnini), volume 3491 of *Lecture Notes in Computer Science (LNCS)*, pp. 271–282. Berlin: Springer.

Leveling, Johannes and Hermann Helbig (2002). A robust natural language interface for access to bibliographic databases. In *Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2002)* (edited by Callaos, Nagib; Maurice Margenstern; and Belkis Sanchez), volume XI, pp. 133–138. International Institute of Informatics and Systemics (IIIS), Orlando, Florida.

Petras, Vivian (2005). GIRT and the use of subject metadata for retrieval. In *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign, CLEF 2004* (edited by Peters, C.; P. D. Clough; G. J. F. Jones; J. Gonzalo; M. Kluck; and B. Magnini), volume 3491 of *Lecture Notes in Computer Science (LNCS)*, pp. 219–225. Berlin: Springer.

Petras, Vivien; Natalia Perelman; and Fredric C. Gey (2004). UC Berkeley at CLEF-2003 – Russian language experiments and domain-specific retrieval. In *Comparative Evaluation of Multilingual Information Access Systems, 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003* (edited by Peters, C.; J. Gonzalo; M. Braschler; and M. Kluck), volume 3237 of *Lecture Notes in Computer Science (LNCS)*, pp. 401–411. Berlin: Springer.

Robertson, Stephen E.; Steve Walker; Susan Jones; Micheline Hancock-Beaulieu; and Mike Gatford (1994). Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)* (edited by Harman, D.), pp. 109–126. National Institute of Standards and Technology (NIST), Special Publication 500-226.