Interpretation of Coordinations, Compound Generation, and Result Fusion for Query Variants

Johannes Leveling School of Computing, CNGL
Dublin City University
Dublin, Ireland
jleveling@computing.dcu.ie

ABSTRACT

We investigate interpreting coordinations (e.g. word sequences connected with coordinating conjunctions such as *"and"* and *"or"*) as logical disjunctions of terms to generate a set of disjunction-free query variants for information retrieval (IR) queries. In addition, so-called hyphen coordinations are resolved by generating full compound forms and rephrasing the original query, e.g. *"rice im- and export"* is transformed into *"rice import and export"*. Query variants are then processed separately and retrieval results are merged using a standard data fusion technique. We evaluate the approach on German standard IR benchmarking data. The results show that: i) Our proposed approach to generate compounds from hyphen coordinations produces the correct results for all test topics. ii) Our proposed heuristics to identify coordinations and generate query variants based on shallow natural language processing (NLP) techniques is highly accurate on the topics and does not rely on parsing or part-of-speech tagging. iii) Using query variants to produce multiple retrieval results and merging the results decreases precision at top ranks. However, in combination with blind relevance feedback (BRF), this approach can show significant improvement over the standard BRF baseline using the original queries.

Categories and Subject Descriptors

H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval—*Query formulation*; H.3.1 [**INFORMATION STORAGE AND RETRIEVAL**]: Content Analysis and Indexing—*Linguistic processing*

Keywords

Query structure, Query variants, Compounds, Result set fusion

1. INTRODUCTION

Information retrieval (IR) approaches typically treat queries as a "bag of words", disregarding any syntactic or logical structure. We investigate a novel approach to generate query variants by interpreting syntactic coordinations of clauses or chunks, indicated by coordinating conjunctions such as *"and"* and *"or"*, as logical disjunctions. In addition, we propose a knowledge-light approach to resolve hyphen coordinations to generate full compounds from short (hyphenated) words used in these coordinations. Query variants are generated by selecting only one conjunct per query variant (subquery). We merge results from processing query variants separately using a standard data fusion technique.

The general idea for the work presented in this paper is to avoid a topic shift when a query puts too much emphasis on all conjuncts, which can occur when a query contains coordinations with many elements or many words. We hypothesize that firstly, resolving hyphenated coordinations generates new compounds that better match actual index terms. For example, in the coordination *"Reisim- und -export"* (rice im- and export), *"im-"* is not a proper word or might be recognized as a stopword. Our proposed approach to compound generation will rewrite this query into *"rice import and export"*. Secondly, processing subqueries focuses on retrieving relevant documents for one aspect of the query. Coordination in queries may shift the focus of retrieved results, either if many less important words are enumerated (e.g. *"Finde Artikel, Reporte, Daten und Dokumente über Bildung"* (Find articles, reports, data, and documents on education)), or when the coordination elements otherwise dominate other key aspects in a query. For example, the query *"Informationen über US-Beziehungen mit Brasilien, Russland, Indien und China"* (Information on US relations with Brazil, Russia, India, and China) would aim for documents describing the relations between the *US* and other countries, but not the relations between *China* and *Brazil*. However, the bag-of-words approach disregards the structural information. Our proposed approach generates four subqueries containing the word pairs {*US, Brazil*}, {*US, Russia*}, {*US, India*}, and {*US, China*}. Results for the subqueries are then merged using a standard result fusion technique to retrieve documents relevant to more than one aspect of the original query (i.e. more than one subquery) at top ranks. Thus, the final data fusion of results retrieved for the query variants would still result in ranking documents covering more than one aspect higher.

The rest of this paper is organized as follows: Section 2 discusses related work. Section 3 introduces our proposed approaches to generate compounds and query variants. Section 4 presents our experimental setup and experimental results. Section 5 presents and analyzes the results, before concluding in Section 6.

2. RELATED WORK

The query understanding workshop at SIGIR 2010[1] is one example of an effort to investigate the structural analysis of IR queries.

Xue and Croft [14] propose to represent queries as sets of distributions, allowing to transform queries, which are associated with

[1] http://ciir.cs.umass.edu/sigir2010/qru/

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

variations generated by applying operations such as adding or replacing words or detecting phrase structures, into a distribution of query reformulations. In comparison, the approaches described in this paper focus on symbolic query manipulation.

Jones et al. [8] investigate query substitutions, a method for generating query suggestions. They propose to replace the original query using information about similar queries extracted from query logs. In contrast, our proposed approach does not rely on additional external query logs or on resources not freely available.

Data fusion of results typically involves multiple unrelated retrieval approaches or retrieval models (see, for example [4]). The idea dates back to Fox and Shaw [5], who conducted experiments on TREC[2] data. Savoy [13] performed extensive experiments on CLEF[3] data and showed that combining results from different retrieval models improves IR performance significantly. In contrast, the experiments in this paper are based on generating subqueries from the same original query and use a single retrieval model.

Compound analysis in IR traditionally focuses on decompounding words and has been shown to improve IR effectiveness for compounding languages such as Finnish, Dutch, or German [1, 2]. For decompounding words, we follow techniques as described by [10, 13], but extend the decomposition procedure to handle additional compound linking elements between compound constituents.

Hyphen coordination has been briefly described by Neumann et al. as part of a shallow NLP engine [11], but it typically requires deeper NLP techniques. Hartrumpf et al. [6] explore question decomposition for geographic questions. Their technique is based on a syntactic-semantic analysis which produces semantic networks. These are split up to generate subquestions, which are then processed separately and answers retrieved for subquestions are used to reformulate the original question.

Huston and Croft [7] explore reducing the length of verbose queries. Our proposed approach to identify coordinated compounds by resolving hyphenated coordinations adds new compounds to a query, but also transforms the original query into simpler and shorter subqueries. The approach proposed in this paper is completely novel and requires – in contrast to the solutions above – no syntactic or semantic parsing or part-of-speech tagging. Only case information, stopwords, and part-of-speech information for closed word categories are employed. Our experiments focus on processing German, but, as the translated examples show, similar problems occur in English and our proposed solution could be applied to other languages as well. The techniques proposed in this paper have – to the best of our knowledge – not been investigated for IR.

3. QUERY PROCESSING

In this section, we briefly describe the approaches proposed to identify and process coordinations. In linguistics, a coordination is a syntactic structure that links together two or more elements (conjuncts). It is usually indicated by a coordinating conjunction (coordinator). The totality of coordinator(s) and conjuncts forming an instance of coordination is called a coordinate structure.

Splitting Compounds. For decomposition of German words, we employ the general process proposed by Koehn and Knight [10]. This process considers each position in a word as a candidate split, where the probability of a split is based on the term frequency of the resulting candidate constituents in a document collection. The decomposition with the maximum probability is selected and the process is applied recursively to the constituents. In contrast to simpler approaches to decompounding (e.g. [3]), which handle only the most frequent linking elements such as *"+s"*, our compound splitter allows a larger set of linking elements and also allows combinations of elisions and linking elements (e.g. *"-e+s"*: *"Mietshaus"* (rental house) = *"Miet(-e)+s+haus"*). The minimum word length for compound constituents is 3. Table 1 shows examples for the linking elements and elisions handled by our compounds splitter.

Table 1: Examples for compound splitting.

Linking	Example Compound	Decomposition
""	*"Bergspitze"* (mountain top)	*"Berg+spitze"*
"+(e)s"	*"Jahresbericht"* (annual report)	*"Jahr+es+bericht"*
"+e"	*"Tagebuch"* (diary)	*"Tag+e+buch"*
"+(e)n"	*"Wolkenbildung"* (cloud formation)	*"Wolke+n+bildung"*
"+(e)r"	*"Kindergarten"* (kindergarten)	*"Kind+er+garten"*
"+(e)ns"	*"Namensraum"* (name space)	*"Name+ns+raum"*
"+nen"	*"Königinnenwitwe"* (queen widow)	*"Königin+nen+witwe"*
"-e"	*"Mieteinnahmen"* (rental income)	*"Miet(-e)+einnahmen"*
"-n"	*"Wartezimmer"* (waiting room)	*"Warte(-n)+zimmer"*
"-en"	*"Rasierapparat"* (razor)	*"Rasier(-en)+apparat"*

Table 2: Frequency distribution of hyphenated coordinations.

Type	Freq.	Example / Expanded Form
1	248,859	*"NATO-Soldaten oder -Flugzeuge"* (NATO soldiers or troops) / *"NATO-Soldaten oder NATO-Flugzeuge"* (NATO soldiers or NATO troops)
2	3,186	*"Öl- und Gasmarkt"* (oil and gas market) / *"Ölmarkt und Gasmarkt"* (oil market and gas market)
3	311	*"Münzzähl- und -verpackungsanlagen"* (coin counting and packaging systems) / *"Münzzählanlagen und Münzverpackungsanlagen"* (coin counting systems and coin packaging systems)

Generating Compounds. In German, common nouns and proper nouns are capitalized, which make them easy to distinguish from adjectives or verbs. For simplicity, we assume that coordinators are expressed as a single word. We employ a small set of coordinators for our experiments (e.g. *"und"* (and), *"oder"* (or), *"sowie"* (as well as), *"/"* (/)). Other coordinators occur only rarely in the topics and are also likely to express contrast or negation (e.g. *"weder ... noch"* (neither ... nor)) and should thus be treated differently.

Queries are tokenized and tokens encompassing a coordinate structure are returned as triples (w_l, w_c, w_r), where w_c is a coordination (e.g. *"and"*, *"or"*) and w_r, the word immediately right of w_c, starts with a hyphen (type 1), or w_l, the word immediately left of w_c ends with a hyphen (type 2), or both (type 3; see Table 2).

The compound generation process works as follows: The word not starting or ending with a hyphen is expected to be a compound and is split into its constituent parts. If this word can not be split, the triple is discarded. A new compound is generated by concatenating the hyphenated form (without hyphen) with the first or last constituent of the compound. For example, *"Reisimport"* (rice import) is split into *"Reis"* (rice) and *"Import"* (import). Concatenating the first constituent with the hyphenated form *"-export"* (export) yields *"Reisexport"* (rice export), The combination of both types (type 3) is possible, but occurs very rarely (e.g. *"Weinan- und -abbau"* (wine growing and harvest)).

As an initial experiment, we analyzed the frequency of hyphen coordinations in a large German corpus of news articles from the Leipzig corpora collection[4]. We used corpus of 16M sentences

[2] http://trec.nist.gov/

[3] http://www.clef-initiative.eu/

[4] http://corpora.uni-leipzig.de/

from news articles from 1995-2010. Table 2 shows the frequency and examples of hyphenated coordinations. 1.5% of all sentences contain hyphen coordination, where 248,859 coordinations are of type 1, 3,186 of type 2, and only 311 of type 3. In contrast, 13 topics out of 300 queries (150 topic titles and descriptions each) in our test data exhibit hyphenated coordinations, all of type 1. As coordinations of type 3 are very rare, we chose to not extend our algorithm to handle this case. Our proposed algorithm handles all 13 cases correctly.

Identifying Coordinations. For simplicity, we do not consider coordination of full sentences or complex phrases as these could express unrelated aspects of a query and seldomly occur in short queries. The main idea of our proposed approach is to identify the coordinate structures and recombine conjunct words to model correct syntactic attachment.

Our proposed approach aims at simplicity and speed and thus, does not rely on parsing or part-of-speech tagging. Let w_1, w_2, ..., w_n be the tokenized text. For each position i where w_i is a coordination, follow the algorithm outlined below.

- Let L be the word sequence left of i, i.e. $w_{start}, ..., w_{i-1}$, which is delimited by the beginning of the text ($start = 0$), a stopword or punctuation at $start-1$, or a case change from upper case to lower case.
- Let R be the word sequence right of i, i.e. $w_{i+1}, ..., w_{end}$, which is delimited by the end of text ($end = n$), or a stopword or punctuation at $end + 1$. Initial articles at position $i + 1$ are skipped.
- Let L_L be the sequence of lowercase words in L, starting at position $start$. L_U is the sequence of uppercase words in L, ending at $i - 1$.
- Let R_L be the sequence of lowercase words in R, starting at position $i + 1$. R_U is the sequence of uppercase words in R, ending at end.
- Combine the sequences L_L, L_U, R_L, and R_U as shown in Table 3 to form a triple (L', w_c, R'). There are 16 possible cases to be considered, of which all but 5 are considered to be incorrect coordinations, corresponding to mismatches in syntactic category or non-constituent conjuncts. For example, if both R_L and R_U are empty sequences, the query remains unchanged.
- To handle coordinations with more than two elements, the triple is then extended by identifying additional coordination elements separated by commas, starting from position $start - 1$ and scanning from right to left. The output is an expanded form of the input, where adjectives are heuristically attached to expand conjuncts.

In the GIRT 2003-2008 topics, there are 123 coordinations, 30 in the topics titles and 93 in the topic descriptions. Topics from 2006 have 23 or more cases per 25 topics (2006: 26, 2007: 23, 2008: 28), topics before 2006 have 17 cases or less (2003: 14, 2005: 17, 2006: 15). Thus, the effect of coordination processing on IR effectiveness should be higher for topics from 2006. A brief post-analysis of the results showed that a few cases in the topics were not handled correctly, due to ambiguity, syntactic complexity (e.g. nested coordinations), or coordination of full sentences.

Generating Query Variants. Query variants are generated by applying (in this order) hyphen coordination resolution, coordination resolution, and compound splitting on a query. For example, the query *"Japans Reisim- und export"* (Japan's rice im- and export) is first transformed into query *"Japans Reisimport und Reisexport"* (Japan's rice import and export) and processed into two query variants *"Japans Reisimport"* (Japan's rice import) and *"Japans Reisexport"* (Japan's rice export). For our experiments, we also add the original queries (title and description) as variants.

Table 3: Resolving coordinations into disjunctions.

Example	Disjunction set
"hohe Intelligenz oder Begabung" (high intelligence or giftedness)	{ *"hohe Intelligenz"* (high intelligence), *"hohe Begabung"* (high giftedness) }
"industrielle Entwicklung und ökonomische Entwicklung" (industrial development and economic development)	{ *"industrielle Entwicklung"* (industrial development), *"ökonomische Entwicklung"* (economic development) }
"Bioprodukte oder ökologische Tierhaltung" (bio product or ecological animal husbandry)	{ *"Bioprodukte"* (bio products), *"ökologische Tierhaltung"* (ecological animal husbandry) }
"Diagnose und Behandlung" (diagnosis and treatment)	{ *"Diagnose"* (diagnosis), *"Behandlung"* (treatment) }
"analysieren oder beschreiben" (analyze or describe)	{ *"analysieren"* (analyze), *"beschreiben"* (describe) }

Merging Results. To merge results retrieved for query variants, we employ the combMNZ approach [5] on document scores from retrieval results $R_1, ..., R_k$, after MinMax normalization. This is a standard data fusion technique used in previous IR research [4, 13].

4. RETRIEVAL EXPERIMENTS

We perform IR experiments on the German data for the GIRT-4 domain-specific task at CLEF (see, for example [9]), because we assumed that coordination occurs frequently in these topics. Our analysis later confirmed this assumption and showed that coordinations occur in 41% of queries generated from title and description (see Table 3). We used data from 2003 to 2008, which comprises 25 topics and their corresponding relevance assessments per year. The GIRT document collection contains more than 150,000 documents from the social sciences.

Queries are preprocessed by case folding, stopword removal, and stemming, using a light German stemmer [13]. Queries are transformed into query variants as described. We used the GIRT topic titles and description fields (TD) as queries. We employ the BM25 retrieval model with default parameters ($k_1 = 1.2$, $b = 0.75$, $k_3 = 1000$). We use standard blind relevance feedback (BRF) [12] with 10 feedback terms and 20 feedback documents, which corresponds to a conservative setting for BRF for this task. The BRF experiment serves as our baseline.

We perform four experiments per topic set, using standard BM25, its combination with BRF, and the corresponding experiments using query variants obtained from interpreting coordinations as disjunctions (QV). In addition, we performed a set of experiments using only queries containing at least one coordinator. We report mean average precision (MAP), GMAP (geometric MAP), the number of retrieved and relevant results (rel_ret), and precision at a cut-off of N (P@N). Results are shown in Table 4. Significance tests are based on the paired Wilcoxon test with 95% confidence level and compare results to the standard BRF to obtain a strong baseline. Significant improvements are indicated by *"*"*.

5. DISCUSSION AND ANALYSIS

Compared to the official results of participants in the GIRT task at CLEF, our results rank among the top three for all years. The better performing systems in this task usually employed a much more complex system setup, e.g. by combining results from different retrieval models, or by using additional knowledge for domain adaptation. In contrast, we did not employ any additional domain

Table 4: Evaluation results for German retrieval on TD fields.

	Parameters		Results			
Data	QV	BRF	MAP	GMAP	rel_ret	P@10
2003	N	N	0.4427	0.3633	1741	0.7280
2003	Y	N	0.4054	0.3319	1681	0.7200
2003	N	Y	**0.4987**	0.3954	1770	0.7120
2003	Y	Y	0.4813 (-3.5%)	0.3754	1789	0.7040
2004	N	N	0.3825	0.2756	1253	0.6000
2004	Y	N	0.3335	0.2263	1224	0.5440
2004	N	Y	**0.4242**	0.2822	1397	0.6320
2004	Y	Y	0.4180 (-1.5%)	0.2775	1406	0.6240
2005	N	N	0.4392	0.3505	2231	0.7440
2005	Y	N	0.4183	0.3242	2160	0.7520
2005	N	Y	0.4585	0.3115	2294	0.7280
2005	Y	Y	**0.4637** (+1.1%)	0.3504	2287	0.7600
2006	N	N	0.4322	0.3649	3080	0.7760
2006	Y	N	0.4047	0.3371	3064	0.8000
2006	N	Y	0.4905	0.3927	3345	0.8160
2006	Y	Y	**0.5009** (+2.1%)	0.4027	3349	0.8000
2007	N	N	0.2892	0.2035	2590	0.5760
2007	Y	N	0.2375	0.1610	2355	0.6080
2007	N	Y	0.3440	0.2351	2975	0.5680
2007	Y	Y	**0.3557*** (+3.4%)	0.2518	3006	0.5920
2008	N	N	0.3482	0.2410	1957	0.6160
2008	Y	N	0.3378	0.2292	1926	0.6480
2008	N	Y	0.3676	0.2078	1985	0.5840
2008	Y	Y	**0.4052*** (+10.2%)	0.2545	2049	0.6240
all	N	N	0.3610	0.2710	8016	0.6345
all	Y	N	0.3288	0.2434	7769	0.6621
all	N	Y	0.3927	0.2644	8481	0.6437
all	Y	Y	**0.4031** (+2.6%)	0.2902	8573	0.6575

adaptation for this domain-specific task, as this paper focuses on processing coordination in queries.

Hyphen coordination is frequent in the topics and occurs in 41% of all tested topic titles and descriptions. Interestingly, the precision and MAP of the initial retrieval run is is typically considerably lower when using query variants compared to the standard initial retrieval. However, the results for the subsequent BRF can outperform results for the standard BRF. This indicates that the generated query variants are better at selecting better (not necessarily relevant) feedback documents. Using the combination of query variants and BRF can produce significantly higher MAP compared to standard BRF, especially when the topics contain many coordinations (e.g. for the 2007 and 2008 topic set).

In addition, coordination might not correspond to logical disjunction in all cases, but this view helps to generate shorter queries. For example, so-called twin pairs such as *"Tag und Nacht"* (night and day), fixed expressions such as *"mehr oder weniger"* (more or less) and short title queries *"Gewalt und Schule"* (violence and schools) should probably not be treated as disjunctions.

6. CONCLUSIONS AND FUTURE WORK

Coordinations are little researched in IR. We propose a simple method to detect coordinations and the corresponding chunks and transform them into query variants. Our results for query variants show a lower precision at top ranks. However, we showed that interpreting coordinations as logical disjunctions in combination with BRF can improve IR performance significantly compared to standard BRF.

The proposed method relies on case information to distinguish between adjectives and nouns which is available in German. To adapt this method to languages which do not capitalize nouns (such as English) or to languages which do not have cases at all (such as Chinese), more complex natural language processing such as part-of-speech tagging is required. The method can then be adapted to work with adjectives (taking the role of lowercase words) and nouns (uppercase words). For English, the GIRT topics showed a similar number of cases with coordinations which illustrates the importance of coordinations in other languages.

As part of future work, we want to apply coordination detection to machine translation. Coordinations can span long distances, which cannot easily be captured by standard MT n-gram models. We expect that reformulating and simplifying the MT input and combining the translation results will increase MT quality.

Acknowledgments

This research is supported by the Science Foundation of Ireland (grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (`http://www.cngl.ie/`).

7. REFERENCES

[1] E. Airio. Word normalization and decompounding in mono- and bilingual IR. *Inf. Retr.*, pages 249–271, 2006.

[2] M. Braschler and B. Ripplinger. How effective is stemming and decompounding for German text retrieval? *Inf. Retr.*, 7(3-4):291–316, 2004.

[3] A. Chen and F. C. Gey. Multilingual information retrieval using machine translation, relevance feedback and decompounding. *Inf. Retr.*, 7(1–2):149–182, 2004.

[4] W. B. Croft. Combining approaches to information retrieval. In *Advances Information Retrieval: Recent Research from the CIIR*, chapter 1, pages 1–36. Kluwer Academic, 2000.

[5] J. A. Fox and E. A. Shaw. Combination of multiple searches. In *TREC-2*, pages 243–252, Gaithersburg, MD, 1994. NIST.

[6] S. Hartrumpf and J. Leveling. Recursive question decomposition for answering complex geographic questions. In *CLEF 2009*, volume 6241 of *LNCS*, pages 310–317. Springer, 2010.

[7] S. Huston and W. B. Croft. Evaluating verbose query processing techniques. In *SIGIR 2010*, pages 291–298. ACM, 2010.

[8] R. Jones, B. Rey, O. Madani, and W. Greiner. Generating query substitutions. In *WWW'06*, pages 387–396, 2006.

[9] M. Kluck. The domain-specific track in CLEF 2004: Overview of the results and remarks on the assessment process. In *CLEF 2004*, volume 3491 of *LNCS*, pages 260–270. Springer, 2005.

[10] P. Koehn and K. Knight. Empirical methods for compound splitting. In *EACL '03*, pages 187–193. ACL, 2003.

[11] G. Neumann and J. Piskorski. A shallow text processing core engine. *Computational Intelligence*, 18(3):451–476, 2002.

[12] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *TREC-3*, pages 109–126, Gaithersburg, MD, 1995. NIST.

[13] J. Savoy. Report on CLEF-2003 monolingual tracks: fusion of probabilistic models for effective monolingual retrieval. In *CLEF 2003*, volume 3237 of *LNCS*, pages 322–336. Springer, 2004.

[14] X. Xue and W. B. Croft. Representing queries as distributions. In *Query representation and understanding workshop at SIGIR 2010*, pages 9–12, 2010.