1 Introduction

The project presented in this thesis explores the phenomenon of word collocation and its application in information retrieval. Specifically the project was aimed at studying long-span collocations and their use in a probabilistic model of IR.

Long-span topic-level collocations are motivated by lexical-semantic relations, as opposed to short-span collocates – motivated by lexical-grammatic or habitual factors Long-span collocates have certain lexical-semantic relatedness, causing their more than random co-occurrence in the same thematic environments – topics – in text. The motivation for studying this group of collocations in their application to IR is to explore ways of obtaining more relevance-discriminating evidence from the contextual environments of query terms in texts.

In IR long-span collocations received relatively little attention, in contrast to short-span collocation, used in phrase identification, or document-wide co-occurrence, used for a wide range of tasks like term weight calculation or relevance feedback. Much of the research on document-wide co-occurrence dates back to the time before the wide use of full-text. However, it is still very widely used, for example in the relevance feedback process of the Robertson & Sparck Jones probabilistic model – the testbed for this project. Accounting for document-wide co-occurrence is reasonable with the abstracts collections, where the contents of an abstract is rather semantically homogeneous. However, it no longer seems adequate for full-text collections, containing multi-topic documents. It was hoped that accounting for term dependencies within smaller and more semantically homogeneous text constructs would lead to improved retrieval performance.

The basic research question of the project is whether the use of long-span collocates of query terms can improve performance of the probabilistic model. The research question was explored by studying three different ways of integrating collocation information into the probabilistic model: global collocation analysis, local collocation analysis and lexical cohesion analysis using local collocations. Four hypotheses, examined over the course of the project, as well as an overview of the above three sets of studies forming the project are given in Chapter 5.

Some of the information presented in sections *5.4 Choice of statistical measures of collocation significance* and *6.3 Lexical-semantic analysis of global collocates* was originally presented in a paper by Vechtomova and Robertson – [Vechtomova2000].

The structure of the thesis is described below.

Chapters 2, 3 and 4 are largely theoretical and contain literature review and discussions.

Chapter 2. *Linguistic perspective on text and lexis* gives a theoretical background of the linguistic aspects of the objects of study. The first section contains a discussion on the organisation of text, what defines a text and what properties differentiate a text from unconnected sequences of sentences. Following [Halliday76] it is argued that a

text is a semantic unit of a non-structural nature, held together by a property of cohesion. A brief description of the hierarchies of linguistic elements forming the plane of expression and the plane of contents of a text is also given. The first section also contains a discussion of the text-forming property of cohesion.

The second section is devoted to lexis as a stratum of the language-system. Specifically, it discusses the notions of a lexical unit, lexical meaning, lexical/semantic relations and syntagmatic/paradigmatic relations between lexical units.

The third section centres around the notion of lexical cohesion and the mechanisms of its realisation in text: lexical links, lexical chains and bonds. A review of the most influential research projects on lexical cohesion in linguistics is also given

The fourth section focuses on linguistic perspectives on collocation. Two different views of collocation are described: collocation as a topic-forming relationship [Halliday76, Hoey91] and collocation due to lexical-grammatical or habitual relations [e.g., Manning99, Palmer81, Robins89]. This section also briefly introduces the branch of linguistics – corpus linguistics – where the methodology for statistical identification of collocation in text was formed. Subsection *2.4.2 Patterns of collocation in text* discusses such aspects of collocation as the distance and consistency of a word's influence on its environment in text.

Chapter 3. *Natural language text in IR context* contains literature reviews and discussions on a number of project-related topics. The chapter starts with a brief discussion of two major approaches to document and query representation in IR: controlled language and natural language indexing.

The second section contains a discussion of two approaches to IR – statistical and natural language processing (NLP). The first half of this discussion is of a more philosophical nature – talking about compatibility of IR with NLP, on the one hand, and with statistical approaches, on the other. The second half compares statistical and NLP approaches to defining complex indexing units (phrases) in IR.

The third section contains literature review of approaches to word collocation in IR, which are divided into three groups: short-span collocations, document-wide co-occurrences and long-span collocations. The later group, being the focus of this thesis, is given the maximum coverage.

The chapter also contains discussions and reviews of IR research on lexical cohesion and document passages.

Chapter 4. *Probabilistic information retrieval* presents the IR model used in this project – the Robertson & Sparck Jones model of probabilistic IR. The chapter focuses on the two major aspects of the model: term relevance weighting and relevance feedback. A separate section talks about term independence assumptions, the model is based on, and a relevance-based dependence, implied by these assumptions.

Chapter 5. *Collocation analysis in this project* starts with the section *5.1 Research question*, containing the formulations of the basic research question and the hypotheses examined in the project. Section *5.2 Overview of the experiments* contains a brief overview of the three stages of experiments, constituting this project: global collocation analysis, local collocation analysis and lexical cohesion analysis using local collocations. These sections were included into chapter 5 instead of the beginning of the thesis, because it was considered important to present the theoretical background first before introducing the experimental part.

This chapter also contains detailed descriptions of the methodologies used in all or most of the experiments in the project, as well as the material used in the experiments: document collection and topics.

Chapters 6, 7 and 8 describe three stages of experiments carried out over the course of the project. A brief overview of these experimental stages is given in chapter 5, section *5.2 Overview of the experiments*.

Chapter 9 *Conclusions and Recommendations* sums up the main findings of the project and presents suggestions for further research.