Expert Systems with Applications 40 (2013) 2530–2540

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Term extraction from sparse, ungrammatical domain-specific documents

Ashwin Ittoo [a,*], Gosse Bouma [b,1]

[a] Department of Operations, Faculty of Economics and Business, University of Groningen, The Netherlands
[b] Computational Linguistics (Information Science), Faculty of Arts, University of Groningen, The Netherlands

ARTICLE INFO

Keywords:
Term extraction
Natural language processing
Text mining
Business intelligence
Product development-customer service

ABSTRACT

Existing term extraction systems have predominantly targeted large and well-written document collections, which provide reliable statistical and linguistic evidence to support term extraction. In this article, we address the term extraction challenges posed by sparse, ungrammatical texts with domain-specific contents, such as customer complaint emails and engineers' repair notes. To this aim, we present ExtTerm, a novel term extraction system. Specifically, as our core innovations, we accurately detect rare (low frequency) terms, overcoming the issue of data sparsity. These rare terms may denote critical events, but they are often missed by extant TE systems. ExtTerm also precisely detects multi-word terms of arbitrarily lengths, e.g. with more than 2 words. This is achieved by exploiting fundamental theoretical notions underlying term formation, and by developing a technique to compute the collocation strength between any number of words. Thus, we address the limitation of existing TE systems, which are primarily designed to identify terms with 2 words. Furthermore, we show that open-domain (general) resources, such as Wikipedia, can be exploited to support domain-specific term extraction. Thus, they can be used to compensate for the unavailability of domain-specific knowledge resources. Our experimental evaluations reveal that ExtTerm outperforms a state-of-the-art baseline in extracting terms from a domain-specific, sparse and ungrammatical real-life text collection.

1. Introduction

The recent years have witnessed a proliferation in unstructured text data. According to several studies (Blumberg & Atre, 2003; Russom, 2007), unstructured texts, in the form of customer complaint emails or engineers' repair notes (e.g. job-sheets), constitute an overwhelming 80% of all corporate data. Buried within these massive amounts of corporate texts are meaningful information nuggets, such as pertinent domain-specific terms. For example, the term "proximity sensor", appearing in engineers' repair notes, indicates that the product component,[2] "proximity sensor", experienced a malfunction, and had to be serviced by an engineer. Efficiently and effectively (accurately) detecting such terms from large repositories of texts is crucial in a wide range of corporate activities. For example, in Product Development-Customer Service (PD-CS) organizations, terms that designate products are useful in "cost of non-quality" analyses for determining which products contribute most significantly to maintenance costs. Terms also provide valuable information on which products fail recurrently and require frequent servicing. This information can be subsequently exploited to improve the product development process, resulting in better quality products and more satisfied customers.

In Natural Language Processing (NLP), several techniques exist for extracting terms from large collections of general and biomedical texts (Ahmad, Davies, Fulford, & Rogers, 1994; Chung, 2003; Dagan & Church, 1994; Frantzi & Ananiadou, 1999; Frantzi, Ananiadou, & Mima, 2000; Uchimoto, Sekine, Murata, Ozaku, & Isahara, 2001). These techniques often rely on external knowledge resources, such as ontologies, which is beneficial for their accuracy (Hiekata, Yamato, & Tsujimoto, 2010; Zhang, Yoshida, & Tang, 2009).

However, most extant term extraction (TE) algorithms are inadequate to address the challenges posed by domain-specific texts, such as those in corporate domains like PD-CS. A major challenge is the sparse nature of these texts, which do not offer reliable statistical evidence, and severely compromise the algorithms' performance. This difficulty is further compounded by the lack of comprehensive domain-specific knowledge resources, for e.g. corporate ontologies, which are difficult to create and to maintain (Auger & Barriere, 2010; Blohm & Cimiano, 2007; Lapata & Lascarides, 2003; Maynard & Ananiadou, 2000; Pecina & Schlesinger, 2006). Another challenge is the detection of multi-word terms, especially those comprising 2 or more words, such as "frequency convertor control board". In addition, there is the issue of ambiguous

* Corresponding author. Address: Nettelbosje 2, 9747 AE Groningen, The Netherlands. Tel.: +31 (0) 50 363 3853.
E-mail addresses: r.a.ittoo@rug.nl (A. Ittoo), g.bouma@rug.nl (G. Bouma).
[1] Address: Oude Kijk in 't Jatstraat 26, 9712 EK Groningen, The Netherlands.
[2] We will use "product" to denote both actual products and their components/parts.

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.10.067

constructs, such as "device is regulating switch" ("*the* device is *the* regulating switch" vs. "*the* device is regulating *the* switch"), and that of incoherent phrases, such as "customer helpdesk collimator shutter". We will elaborate on these difficulties in Section 2.4.

In response to the above challenges and to the growing need of organizations for extracting terms from corporate texts, we present ExtTerm, a novel framework for domain-specific TE. Our core contributions are as follows:

- ExtTerm accurately detects terms from sparse, domain-specific text collections that do not offer reliable statistical evidence, thereby overcoming the issue of data sparsity.
- We extract arbitrarily long terms, including those containing more than 2 words.
- Our term extraction approach is unsupervised, eschewing the need for domain-specific knowledge resources (e.g. ontologies), which are not always available and expensive to construct manually. Instead, we only rely on a readily-available resource, viz. Wikipedia, as a knowledge base. This also shows that readily-available resources, such as Wikipedia, despite their general contents, can still be exploited for domain-specific TE. Thus, they can be exploited to compensate for the lack of domain-specific resources.
- ExtTerm accurately discriminates between valid terms and other ambiguous and incoherent expressions. Many of the latter expressions tend to exhibit some of the core properties of terms, and are thus incorrectly extracted by most existing term extraction systems.

In our experiments, we evaluate the performance of ExtTerm over a real-life, domain-specific text collection, which was provided by our industrial partners.[3] The results reveal that ExtTerm achieves a very high accuracy level in domain-specific TE, and even outperforms the state-of-the-art algorithm of Frantzi et al. (2000).

This article is organized as follows. In Section 2, we present some basic notions associated with terms, describe existing work on automatic term extraction, and highlight the challenges posed by domain-specific, sparse, and informally-written texts. In Section 3, we develop our proposed methodology for term extraction from domain-specific documents. Experimental evaluations and performance comparisons against baselines are given in Section 4. We conclude and discuss areas of future work in Section 5.

In the remainder of this article, we refer to multi-word terms, for e.g. "proximity sensor", as complex terms, and to single-word terms, for e.g. "footswitch", as simple terms. We will also use "text collection", "texts" and "corpus" (plural: "corpora") interchangeably.

2. Preliminaries and related work

2.1. (Domain-specific) terms vs. (general) words

Terms are formally defined as lexical manifestations of domain-specific concepts. Thus, unlike general words, such as "Friday meeting", terms convey a particular meaning in a given domain. For instance, the term "frequency converter control box" designates a specific product in the domain of Product Development-Customer Service (PD-CS) organizations.

At the surface level, however, terms and general words are indistinguishable from each other since they are both realized as strings. Furthermore, as shown in past studies in terminology, terms tend to adopt all the word formation rules in a language (Sager, 1998).

Two properties that enable us to discriminate between terms and general words are the unithood and termhood.

2.2. Properties of terms: unithood and termhood

2.2.1. Unithood

Unithood determines whether an expression is well-formed and behaves as a coherent, atomic linguistic unit (Pazienza, Pennacchiotti, & Zanzotto, 2005). A well-formed expression adheres to a certain syntactic structure, such as noun phrases, for e.g. "frequency converter control box". An expression is a coherent and atomic unit if its individual words tend to co-occur (together) more often than spuriously, i.e. the words are strong collocated. For example, "frequency converter control box" and "proximity sensor" are coherent units, while "customer amplification" is not. Based on its definition, the unithood property is applicable only to complex terms. Simple terms always have perfect unithood since they consist of only 1 word (Kageura & Umino, 1996).

2.2.2. Termhood

Termhood determines whether an expression is representative of a domain. For example, "frequency converter control box" is representative of the PD-CS domain, while "Friday meeting" is not.

2.3. Automatic term extraction approaches

Natural Language Processing (NLP) techniques for term extraction (TE) are based on 3 main approaches, namely Linguistic, Statistical and Hybrid, as will be briefly described below. For a more comprehensive review on the techniques, we refer the reader to the work of Pazienza et al. (2005).

2.3.1. Linguistic techniques

Linguistic techniques operate upon the premise that since terms denote specific concepts, they can only be realized by certain word types (Wright & Budin, 1997). Linguistic techniques are often implemented as Part-of-Speech (POS) filters, such as that of Justeson and Katz (1995), which accepts, as terms, any noun sequences containing optional adjectives and/or prepositions. We will revisit this filter in our experiments (Section 4). Since linguistic techniques rely on the syntactic structure, they identify terms according to the unithood property.

2.3.2. Statistical techniques

Statistical techniques, such as those of Pecina and Schlesinger (2006), estimate the unithood of 2-word expressions, for e.g. "proximity sensor", by computing the collocation strength between the word pairs. This is achieved using well-known Lexical Association Measures (LAM), such as mutual information (Church & Hanks, 1990). Petrovic, Snajder, and Basic (2010) extended several traditional LAMs for estimating the unithood of expressions with at most 4 words.

Concerning the termhood, it is statistically determined based on the observation that the highly frequent expressions in a domain-specific corpus are likely to denote relevant terms. Popular statistical techniques for termhood estimation include the "term frequency–inverse document frequency" (tf–idf) (Salton, 1991) and the C-value (Frantzi et al., 2000).

Another termhood estimation technique is that of corpus comparison, in which a domain-specific corpus is compared against a collection of general texts. Expressions that are more likely in the domain-specific corpus are then treated as domain-specific terms. Several corpus comparison techniques for term extraction are mentioned in literature, such as those of Ahmad et al. (1994), Chung (2003), Drouin (2003), Rayson and Garside (2000) and Romero, Moreo, Castro, and Zurita (2012).

[3] Multi-national corporations based in The Netherlands.