Chapter 9

Mining Multiword Terms from Wikipedia

Silvana Hartmann
Technische Universität Darmstadt, Germany

György Szarvas
Technische Universität Darmstadt, Germany & Research Group on Artificial Intelligence,
Hungarian Academy of Sciences, Hungary

Iryna Gurevych
Technische Universität Darmstadt, Germany

ABSTRACT

The collection of the specialized vocabulary of a particular domain (terminology) is an important initial step of creating formalized domain knowledge representations (ontologies). Terminology Extraction (TE) aims at automating this process by collecting the relevant domain vocabulary from existing lexical resources or collections of domain texts. In this chapter, the authors address the extraction of multiword terminology, as multiword terms are very frequent in terminology but typically poorly represented in standard lexical resources. They present their method for mining multiword terminology from Wikipedia and the freely available terminology resource that they extracted using the presented method. Terminology extraction based on Wikipedia exploits the advantages of a huge multilingual, domain-transcending knowledge source and large scale structural information that can identify potential multiword units without the need for linguistic processing tools. Thus, while evaluated in English, the proposed method is basically applicable to all languages in Wikipedia.

INTRODUCTION

Automated ontology construction, or ontology learning, has received substantial research interest in recent years, as the manual development of formal knowledge models is labor-intensive and cannot scale up to practical needs in the Semantic Web. Terminology extraction—i.e., the automated collection of domain terminology—is the first step towards computer-assisted ontology construction (Cimiano, 2006).

The terminology of a domain (referred to as terms) consists of a subset of general-language lexical units that have a domain-relevant mean-
Mining Multiword Terms from Wikipedia

... and lexical units of the domain-specific sublanguage—i.e., technical terms. Accordingly, terminology extraction aims at finding domain-specific and general domain-relevant lexical units, where the particular domain is defined by the actual application. Figure 1 presents the continuum of domain specificity of lexical units, ranging from general-language units to specialized technical terms (Cabré, 1999). Multiword expressions are interpreted as lexical units which consist of several words and whose irregular semantic, syntactic, pragmatic or statistical properties justify their own entry in a natural-language lexicon (Sag, Baldwin, Bond, Copestake, & Flickinger, 2002). In this chapter, we will refer to domain-relevant multiword expressions as multiword terms.

Typically, the majority of domain-specific vocabulary consists of multiword terms (Nakagawa & Mori, 1998), which makes the extraction of multiword terminology an important problem on its own. In this chapter, we focus on the automatic extraction of multiword terminology, as multiword units (particularly domain-specific ones) are poorly represented in standard lexical resources like WordNet (Sag, et al., 2002). Since ontology construction might address any particular domain, or even domain-transcending areas such as e-learning, we aim at the extraction of a general-purpose multiword lexicon, which can later be filtered according to the particular application needs. We consider our resource to be a first step towards creating parameterized terminology resources, which allows flexible term selection for efficient ontology construction on the fly. A demand for such resources emerged as a consequence of advances in semi-automatic ontology construction and increasing employment of ontologies in semantically enhanced applications. In this context, Wikipedia is an ideal source for terminology extraction, due to its good coverage of a wide variety of domains in multiple languages and its encyclopedic style, placing an emphasis on specialized vocabulary, rather than expressions of linguistic interest, such as idioms.

The proposed flexible terminology resources require dynamic domain adaptation—i.e., the selection of terms for a particular application domain. Domain adaptation typically happens

Figure 1. Properties of terms: term size vs. degree of domain specialization