Semi-Automatic Ontology Development:

Processes and Resources

Maria Teresa Pazienza
University of Roma Tor Vergata, Italy

Armando Stellato
University of Roma Tor Vergata, Italy

Information Science
REFERENCE

Managing Director:	Lindsay Johnston
Senior Editorial Director:	Heather Probst
Book Production Manager:	Sean Woznicki
Development Manager:	Joel Gamon
Development Editor:	Michael Killian
Acquisitions Editor:	Erika Gallagher
Typesetter:	Milan Vracarich, Jr.
Cover Design:	Nick Newcomer, Lisandro Gonzalez

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Semi-automatic ontology development : processes and resources / Maria Teresa Pazienza and Armando Stellato, editors.
p. cm.
Includes bibliographical references and index.
Summary: “This book includes state-of-the-art research results aimed at the automation of ontology development processes and the reuse of external resources becoming a reality, thus being of interest for a wide and diversified community of users”-- Provided by publisher.
ISBN 978-1-4666-0188-8 (hardcover) -- ISBN 978-1-4666-0189-5 (ebook) -- ISBN 978-1-4666-0190-1 (print & perpetual access) 1. Semantic Web. 2. Ontologies (Information retrieval) 3. Semantic networks (Information theory) 4. Automatic data collection systems. 5. Knowledge acquisition (Expert systems) I. Pazienza, Maria Teresa. II. Stellato, Armando, 1975-
TK5105.88815.S468 2012
006.3’3--dc23
2011051816

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

All work contributed to this book is new, previously-unpublished material. The views expressed in this book are those of the authors, but not necessarily of the publisher.

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-

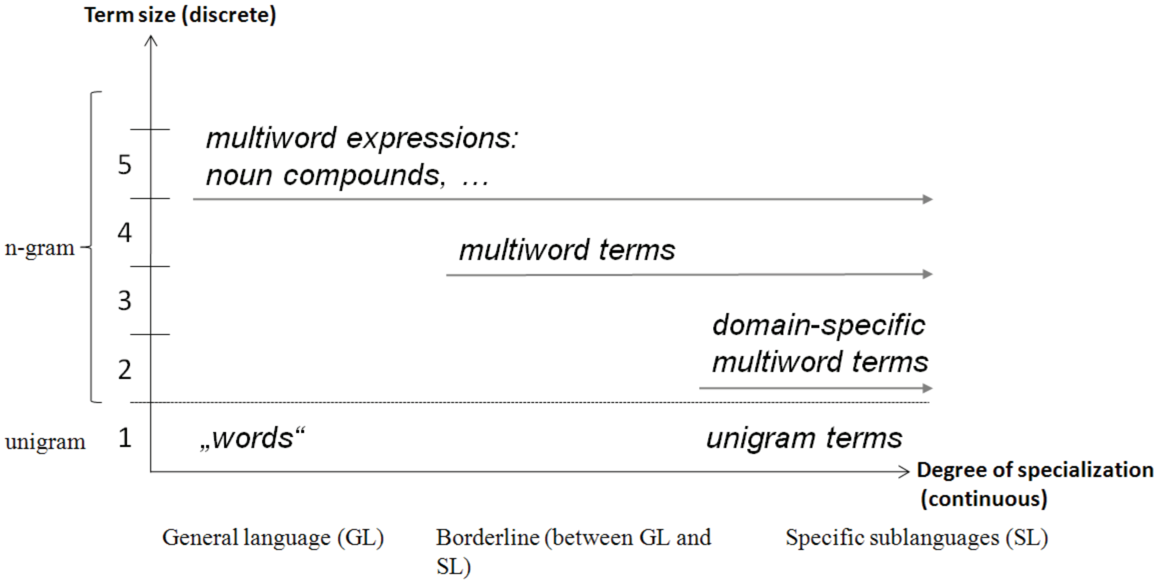
DOI: 10.4018/978-1-4666-0188-8.ch009

ing, 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

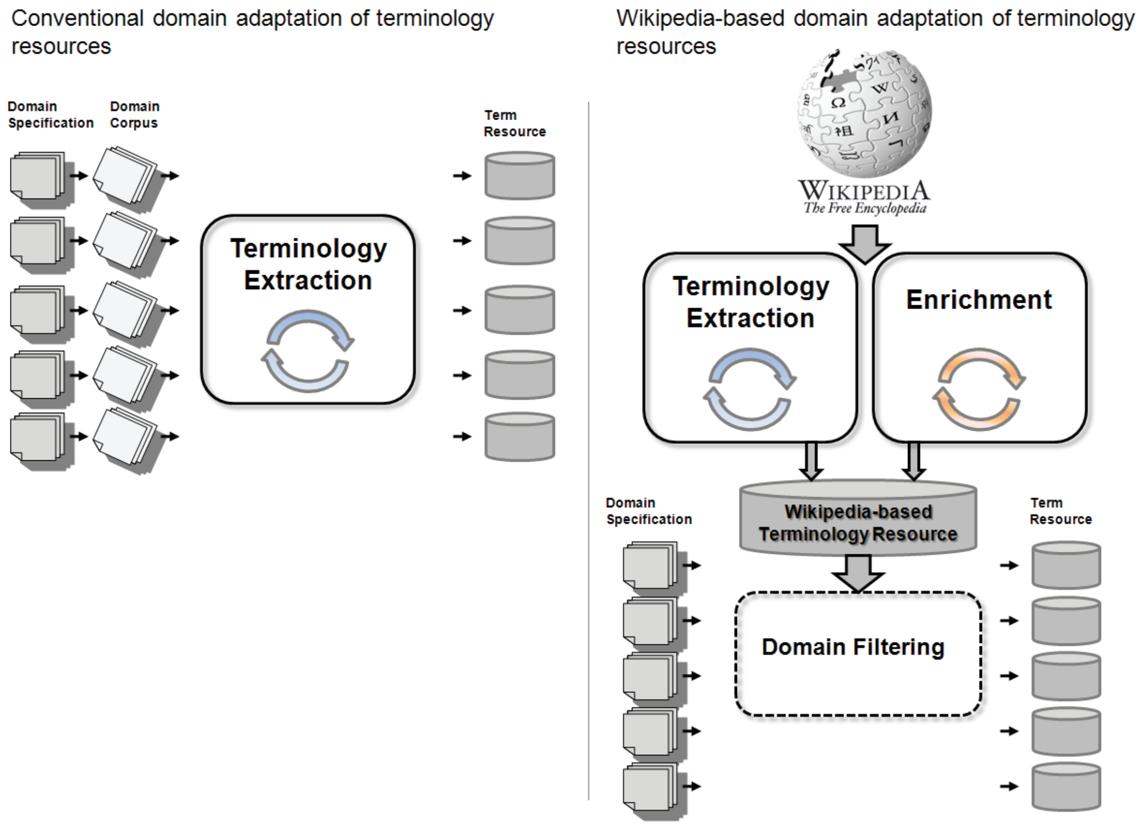
Figure 1. Properties of terms: term size vs. degree of domain specialization


Term size (discrete)
n-gram
unigram
5
4
3
2
1
multiword expressions: noun compounds, ...
multiword terms
domain-specific multiword terms
„words“
unigram terms
Degree of specialization (continuous)
General language (GL)
Borderline (between GL and SL)
Specific sublanguages (SL)

in the corpus collection stage of the terminology extraction cycle: for every new domain, a corpus of domain texts containing the domain-relevant terms is collected. Alternatively, we suggest performing domain adaptation as *domain filtering* on the Wikipedia-based terminology resource independent of the terminology extraction step. Our approach enables ad-hoc building of terminology resources for different domains and degrees of language specialization, and thus improves the lifecycle of terminology building: instead of running through the term extraction process—from corpus collection to term selection—for every new terminology resource, the term extraction process is run only once on Wikipedia. Then the term selection is performed on the Wikipedia-based resource for any domain. Figure 2 illustrates the difference between conventional domain adaptation and enhanced domain adaptation on the Wikipedia-based resource. Although we do not perform the domain filtering ourselves in this work, we suggest ways how it can be done based on the information contained in our resource.

In this chapter, we present and evaluate the extraction process of our terminology resource and its enrichment with category and definition information from Wikipedia—information which can be used in the further ontology construction process. To the best of our knowledge, the present work is the first to evaluate Wikipedia as a source of multiword terms (other than named entities). Related work (Erdmann, Nakayama, Hara, & Nishio, 2008; Erdmann, Nakayama, Hara, & Nishio, 2009) exploits Wikipedia for *bilingual*

Figure 2. Difference between conventional and enhanced, Wikipedia-based domain adaptation of terminology resources


Conventional domain adaptation of terminology resources
Domain Specification
Domain Corpus
Term Resource
Terminology Extraction
Wikipedia-based domain adaptation of terminology resources
WIKIPEDIA The Free Encyclopedia
Terminology Extraction
Enrichment
Wikipedia-based Terminology Resource
Domain Specification
Term Resource
Domain Filtering

terminology extraction of unigram terms and multiword terms. They, however, evaluate their approach only on pairs of unigram terms, but not on the extracted multiword terms.

The proposed resource of multiword terms from Wikipedia is made publicly available to the research community; thus it can be evaluated in specific applications and serve as a base model for further development of flexible terminological resources for semi-automatic ontology construction.

Exploiting the unique characteristics of Wikipedia as a knowledge source offers the following advantages over terminology extraction from domain corpora:

- The approach is in general language-independent, since it does not rely on linguistic text analysis. Many previous approaches that extract terminology from domain-specific corpora use, for instance, Part-of-Speech (POS) patterns or syntactic parses. The absence of robust analysis tools for certain domains or languages might prohibit the application of such methods.
- Wikipedia provides high-quality multiword term candidates. Using Wikipedia as a source for the extraction of multiword terms, we rely on phrase boundaries explicitly marked by humans—i.e., we accept only those phrases as candidates which are explicitly highlighted by different typesetting (bold, italics) or wiki markup (links, link anchor texts, titles, headers). As a result, the extracted multiword term candidates are less noisy than those extracted from general texts with a knowledge-poor approach (e.g., n-grams).
- Wikipedia is a good source of domain-relevant terms: Wikipedia's broad coverage of various specialized domains and its quick evolution with respect to coverage of newly emerging scientific or technological areas makes it a uniquely well-suited resource for terminology extraction to support the construction of formal ontologies in new areas. Thus, Wikipedia is an attractive alternative to the collection of domain-specific texts for terminology extraction.

We note here that, even though the proposed method does not inherently rely on domain-specific texts or complex linguistic analysis, we can naturally exploit these when they are available: we might make use of domain-specific texts and/or part-of-speech information to further filter the extracted candidate lists. Particularly in our study, we will make use of a part-of-speech tagger and a named entity tagger, as for English these tools are easy to obtain. Still, an important aspect of our method is that the use of such tools is not mandatory.

In the following sections, we first provide an overview of the state-of-the-art approaches to 1) term extraction—specifically, related work on term extraction for ontology construction, 2) multiword expression extraction, and 3) using knowledge extracted from Wikipedia in semi-automatic ontology construction. A particular focus is on extracting multiword terminology as opposed to unigram terms, also called *simple terms*. We also introduce Wikipedia and the various types of information contained therein.

In the main part of the chapter, we present our work on extracting multiword terminology from Wikipedia. Our analysis shows that over one million multiword term candidates consisting of two to four words can be extracted from the English version of Wikipedia using the method presented in the chapter. However, not all of the marked-up phrases are valid multiword terms; some of them are conventional natural language phrases, such as "list of countries." Therefore, the candidate phrases identified from Wikipedia are ranked by a statistical measure used in multiword expression mining which exploits corpus statistics of the multiword units and their constituent terms. Based on the ranking, the top-ranked phrases are selected as multiword terms. We describe the steps of this process ranging from candidate extraction and candidate ranking to the final filtering step separating named entities from multiword terms. The extracted multiword term resource is further augmented with definitions and category information from Wikipedia. For evaluation, a sample of the extracted multiword terms is evaluated by human raters. Additionally, we present a comparison of the resource to general-domain multiword terms represented in the Princeton WordNet (Fellbaum, 1998). The chapter closes with a discussion of future research directions and a summary of the presented work.

BACKGROUND

Terminology Extraction

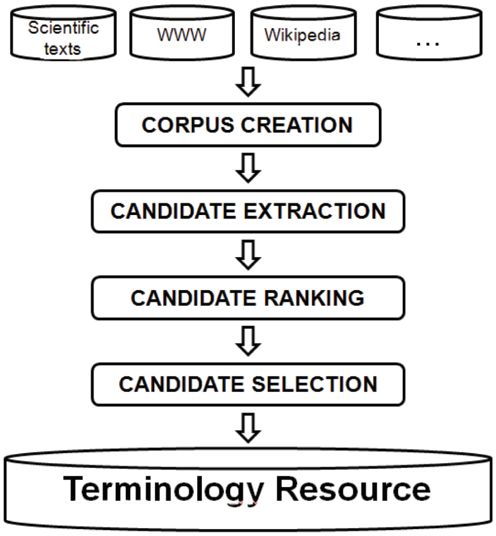
Defining the terminology of a domain is a basic yet laborious task in ontology construction—particularly if performed manually by human experts. As a result, there is a high demand for automated solutions based on natural language processing to support this time consuming and costly process. In automated terminology extraction, domain-relevant terms are mined from text collections exploiting linguistic properties of terms, such as typical phrase structure patterns, their statistical distribution in corpora, or idiosyncratic properties in a particular domain (as with protein names in molecular biology).

The extracted terms serve as input to the later steps of the ontology construction process. The final composition of a vocabulary of terms depends on the type of ontology to be developed: task ontologies (e.g., travel booking as in Gómez-Pérez, Fernández-López, and Corcho, 2004) require a detailed description of events and general world knowledge, while formal domain ontologies often require highly specialized knowledge and scientific terminology. Scientific terminology is very productive—new terms are created continuously. Therefore, techniques for automatic terminology extraction from texts, also called automatic term recognition, are required to efficiently create and maintain terminological resources.

Figure 3 introduces the architecture of the terminology extraction process. It starts with the collection of a corpus representing the target domain. From this corpus, term candidates are extracted and ranked according to their domain relevance. A subset of the ranked candidates is then selected to build the terminology resource. We describe each of these steps of the terminology extraction process in the following paragraphs.

Corpus creation. Corpora for terminology extraction are usually created from collections of domain-specific texts. Such collections can be obtained from edited publications—e.g., technical documentation (Aussenac-Gilles, Biébow, & Szulman, 2000)—or crawled from the web using targeted web search queries (Brunzel, 2008). The former approach yields high-quality texts, but access to large amounts of text might be problematic for certain specialized (or newly emerging) domains. The latter approach poses the problem of data quality management both on the surface

Figure 3. Terminology extraction architecture



level (HTML cleaning, boilerplate removal, etc.) and on the content level (texts of low or questionable quality are common in the Web 2.0). These problems can be avoided by relying on high-quality, easily accessible, yet large-scale sources. Thus, the collaborative encyclopedia Wikipedia, which has proven to be of high quality with respect to text editing and information content (Giles, 2005), has been identified as an information source for corpus construction. Cui, Lu, Li, and Chen (2008), for instance, propose a method for automatically extracting domain corpora from Wikipedia.

Candidate extraction. Candidate extraction techniques using linguistic information exploit the fact that domain-specific terms are typically noun phrases. Conventional approaches extract noun phrases from automatically POS-tagged texts using manually defined regular expression patterns on POS tags.

For example, Frantzi, Ananiadou, and Mima (2000) use the patterns (Noun Noun+), (Adj* Noun+), and (((Adj|Noun)+ |((Adj|Noun)*(Noun Prep)?)(Adj|Noun)*)Noun) to cover simple terms and noun compounds of variable length (e.g., "peptide," "signal peptide"), sequences of adjectives of variable size followed by at least one noun (e.g., "gross national product") and more complex terms comprised of sequences of adjectives, nouns and prepositions (e.g., "language acquisition in children"). The first pattern, retrieving only noun compounds, is more restrictive than the other patterns. This leads to higher precision, since noun compounds have a high likelihood of being domain terms, but lower recall, since terms containing adjectives and prepositions are not found. The patterns can be adapted to the requirements of specific domains regarding the *size* of term candidates—i.e., the number of words they contain, and their internal structure (for instance, whether they include other multiword terms).

If robust linguistic processing is not available, knowledge-poor approaches to term candidate extraction can be applied. The simplest one is to extract n-grams—i.e., continuous sequences of *n* words, from texts. This yields term candidates up to a pre-defined size *n*. Often stop words, such as function words like articles or auxiliary verbs, are filtered out before extracting the n-grams to restrict the size of the candidate set. Since this technique does not take the linguistic phrase structure into account, the mined term candidates are often noisy; they may, for instance, violate phrase structure constraints. Thus, this approach relies heavily on the subsequent candidate ranking step to identify high-quality terms. Moreover, the ranking of all n-grams up to a certain size of *n* might be computationally expensive.

To summarize, the linguistically informed approach requires more resources—namely, the availability of a POS tagger with robust performance in the target domain. The purely statistical approach can operate without it but consequently yields lower precision.

Another very different approach to candidate extraction is to exploit specific properties of a particular text source. While texts extracted from the web often pose difficulties for linguistic processing tools due to low text quality (which may be inherent to the texts or caused by removal of HTML markup), they contain structural information which can be used to identify term candidates: in light of this, Brunzel (2008) uses the XHTML markup in web texts to identify term candidates. XHTML tags, such as headers or emphasis tags, are used to identify suitable candidate sequences. Similarly, the MediaWiki markup in Wikipedia, highlighting Wikipedia article titles and link anchors, has been used to identify candidates for named entity recognition (Toral & Muñoz, 2006) and bilingual terminology extraction (Erdmann, et al., 2008). This approach exploits a higher degree of knowledge on phrase boundaries, since the marked-up sections are typically created by human editors. Still, the highlighted sections are not selected with terminology extraction in mind. This makes a ranking and filtering step necessary.

Candidate ranking and selection. Depending on the characteristics of the text collection and the candidate extraction method applied, a large number of term candidates are extracted that are subsequently ranked and filtered according to their domain-relevance. The final terminological resource is then selected from the ranked list. Either a predefined number of top-ranked terms are retained, or a cutoff value of the ranking function is determined, maximizing the accuracy on a predefined set of gold-standard terms.

The ranking process has to take into account two different characteristics of the term candidates. For simple terms, the property of *termhood*—i.e., whether the candidate represents a domain-specific concept—has to be measured. For multiword terms, the property of *unithood* has to be measured additionally. This property refers to the degree to which the words in a term are associated with each other to form a lexical unit (Kageura & Umino, 1996).

Valid lexical units (i.e., those with a high unithood value) are phrases that should obtain a separate entry in the lexicon ("gross national product"), while phrases that consist of an ad-hoc combination of terms and can be interpreted compositionally (such as "national product of the EU states") indicate a low unithood value. A large number of statistical methods based on the candidates' (and optionally, their constituents') frequency in a domain-specific corpus have been suggested for ranking. Some methods also consider the textual contexts of the terms in a corpus, or the distribution of the terms in another corpus from a different domain. Among the widely used methods are measures that model either the unithood, or the termhood of the candidates, and *hybrid methods* that try to combine both types of measures.

Figure 4 summarizes the most important term ranking approaches based on corpus statistics, grouped by whether they target termhood, unithood or both. Methods that measure unithood—also referred to as *statistical association measures*—are targeted to multiword terms and cannot be applied to unigram terms. They evaluate either the strength of the association between the words in a multiword term candidate like Pointwise Mutual Information (PMI), or measure the significance of this association like χ^2 or Log-Likelihood Ratio (LLR) (Krenn & Evert, 2001). Some methods for the ranking of multiword terms can naturally deal with terms of varied length (see column n-grams in Figure 4), while a few of them are optimized for bigrams and have to be adapted to longer terms (Da Silva & Lopes, 1999). Termhood and hybrid methods, on the other hand, can be used to rank both unigram and multiword terms alike, and can naturally deal with terms of varied length (see column n-grams in Figure 4).

We further distinguish between ranking methods that use frequency information about the term only, and methods that also consider the textual context around the terms (column *Corpus Context* in Figure 4). Term frequency in a domain-specific corpus has also been used as a (baseline) ranking method. Frequency, however, is biased towards short terms, as unigrams have higher frequencies in a corpus than n-grams. Thus a size factor penalizing shorter term candidates is usually included in more sophisticated methods (such as the lexical cohesion measure in TermExtractor and Glossex—see Figure 4). In a similar manner, unithood methods and hybrid methods take into account the frequencies of both the term candidates and their constituent words. Tf-idf, a weighting scheme originally introduced in information retrieval (Salton, Wong, & Yang, 1975), relates the frequency of a term in a document to the number of documents the term occurs in (document frequency). The value is higher, if a term occurs frequently in a document, but only in few documents across the full corpus. Terms with a high tf-idf value are considered domain-relevant.

The C-value and NC-value (Frantzi & Ananiadou, 1999) also take contextual information into account. C-value considers a special form of contextual information: the occurrence of a term candidate c as part of a longer term e, called

Figure 4. Overview of statistical methods for term extraction

	Method	Type	n-grams	Corpus Context	Contrast Corpus	Formula
1	**Frequency**	termhood	1+	-	-	$f(c, D)$
2	**Tf–idf**		1+	-	-	$\text{tf–idf}(c) = f(c, D) \cdot \frac{\lvert D \rvert}{df(c)}$
3	**Weirdness**		1+	-	+	$W(c) = \frac{f(c,D)/t(D)}{f(c,C)/t(C)}$
4	**PMI**	unithood	2	-	-	$PMI(c = w_1 w_2) = \log \frac{P(c,D)}{P(w_1,D) \cdot P(w_2,D)}$
5	**Log-likelihood ratio**		2	-	-	$LLR(c) = -2 \sum_{ij} f(ij) \log \frac{f(ij)}{\hat{f}(ij)}$
6	**Pearson's χ^2**		2	-	-	$\chi^2(c) = \sum_{i,j} \frac{(f(ij) - \hat{f}(ij))^2}{\hat{f}(ij)}$
7	**C-value**	hybrid	1+	+	-	$C(c) = \begin{cases} \log_2 n \cdot f(c, D) & \text{- if c is not nested} \\ \log_2 n \cdot (f(c, D) - \frac{1}{\lvert N_c \rvert} \sum_{e_i \in N_c} f(e_i, D)) & \text{- otherwise} \end{cases}$
8	**NC-value**		1+	+	-	$NC(c) = 0.8 \cdot C(c) + 0.2 \cdot \sum_{m \in M} f(c, m, D) \cdot \omega(m, D)$
9	**Glossex**		1+	-	+	$GL(c) = \alpha \cdot TD(c) + \beta \cdot TC(c)$ Domain Specificity: $TD(c) = \frac{1}{n} \cdot \sum_{w_i \in c} \log \frac{P(w_i,D)}{P(w_i,C)}$ Term Cohesion: $TC(c) = \frac{n \cdot f(c,D) \cdot \log(f(c,D))}{\sum_{w_i \in c} f(w_i,D)}$
10	**TermExtractor**		1+	-	+	$\text{TermExtractor}(c) = \alpha \cdot DR(c) + \beta \cdot DC(c) + \gamma \cdot LC(c)$ Domain Relevance: $DR(c) = \frac{f(c,D)}{\max_j f(c, c_j \in C)}$ Domain Consensus: $DC(c) = -\sum_{d_i \in D} P(c, d_k) \log(P(c, d_k))$ Lexical Cohesion: $LC(c) = TC(c)$

Legend:

- Term candidate $c = w_1, \ldots, w_n, e$; number of words w in term c: $|c|$
- Domain corpus D, contrast corpus C; documents $d \in D, c \in C$; # of words in corpus X: $t(X)$
- Frequency of term c in X: $f(c, X)$; document frequency of term c in corpus X: $df(c, X)$
- Number of documents in corpus X: $|X|$
- Candidate term e contains term c (nestedness); $N_c = e_i, \ldots, e_n$ set of nested terms, $|N_c|$ = # of nested terms
- $M = m_1, \ldots, m_k$ set of "marker words"; $\omega(m, D)$: weight of marker word determined from D
- Frequency of c in X next to marker word m: f(c,m,X)
- Relative frequency of w in X: $P(w, X) = f(w,X)/t(X)$
- Parameters $\alpha, \beta, \gamma, \ldots$
- Contingency table: observed probabilities:

	w_2	$\cancel{w}_2$	
w_1	f_{11}	f_{12}	$f(w_2)$
$\cancel{w}_1$	f_{21}	f_{22}	$f(\cancel{w}_1)$
	$f(w_2)$	$f(\cancel{w}_2)$	

- In contingency table: $\hat{f}(ij)$ – expected frequency, under independence assumption: $\hat{f}(12) = (f(w_1)f(w_2))/N$

"nestedness." The enclosed occurrences f(e, D) are subtracted from the frequency of the term candidate, as they are not counted as evidence of the enclosed term. Thus the occurrence of "national product" nested in "gross national product" is not counted. Based on the observation that enclosed terms which occur in a large variety of contexts also occur independently, the number of different contexts the candidate appears in $|N_c|$ is used to normalize the subtracted number, as shown in Figure 4.

NC-value is a weighted sum of C-value and a "context information factor" (Frantzi & Ananiadou, 1999). The context information factor quantifies the assumption that there are specific words in a domain that frequently co-occur with domain-specific terms. These words are interpreted as markers of termhood. Frantzi and Ananiadou (1999) identify a set of these markers using a seed set of manually selected terms. The most frequent content words (nouns, verbs, or adjectives) occurring directly before or after the seed terms in a domain-specific corpus are selected as termhood markers. These markers receive a weight—ω(m, D) in Figure 4—based on the frequency they are found together with a term. The context information factor sums over the number of occurrences of the candidate next to the marker multiplied with the weight of the marker as shown in Figure 4.

Weirdness, TermExtractor, and Glossex use additional corpora besides those the candidates are extracted from. This is shown in column *Contrast Corpus* in Figure 4.

Weirdness (Ahmad, Gillam, & Tostevin, 1999) compares the relative term frequencies in the domain corpus to those in a general newswire corpus. Terms which have high weirdness are more closely related to the target domain.

Glossex (Kozakov, et al., 2004) incorporates a general-domain corpus as part of a domain specificity measure TC(c), which consists of the average log weirdness of a term's constituent words.

TermExtractor (Sclano & Velardi, 2007) is similar to Glossex, but uses a set of out-of-domain corpora, (i.e., domain-specific corpora from domains other than the target domain) to compute domain relevance DR(c) instead: this measure compares the frequency of a term candidate in the domain corpus to the highest frequency in the set of out-of-domain corpora.

Evaluation. To evaluate a particular term extraction method, the extracted terms are usually either compared to a terminological dictionary, or the top-ranked terms are manually annotated for domain relevance by a group of domain experts. Depending on the chosen evaluation strategy, precision (i.e., the proportion of correct terms in the list of extracted terms) and/or recall (i.e., the proportion of terms retrieved to the complete set of terms in the corpus) of the studied methods—and, of course, variants of these measures—can be estimated.

Both evaluation strategies have advantages and disadvantages: recall can only be measured with respect to an existing terminology resource, which is often not available in sufficient quality and size. This evaluation strategy, furthermore, does not consider that a term extraction method is able to extract previously unknown terms—exactly what it is required to do—and therefore may underestimate precision. Precision can be more reliably estimated by manually rating the extracted terms. As manual annotation is time consuming, typically only a subset of the extracted terms can be evaluated.

Comparison of term extraction techniques. Which method performs best in terminology extraction is essentially an open research question. Several works in the recent years compared different term extraction techniques on various domain corpora (Pazienza, Pennacchiotti, & Zanzotto, 2005; Korkontzelos, Klapaftis, & Manandhar, 2008; Zhang, Iria, Brewster, & Ciravegna, 2008). All these studies compare several popular methods of term extraction. They aim at identifying the best method based on an evaluation under the same conditions.

The term extraction methods compared include 1) frequency, C-value, NC-value, PMI, and significance of association measures (t-test, χ^2, LLR) (Korkontzelos, et al., 2008), 2) degree of association measures (mutual information), significance of association measures (t-test, LLR), frequency and C-value (Pazienza, et al., 2005), and 3) tf-idf, weirdness, C-Value, re-implementations of TermExtractor and Glossex (Zhang, et al., 2008). Note that the third set includes only those methods that can be applied to both simple and multiword terms, as Zhang et al. (2008) propose an integrated approach for both types of terms.

The studies mainly target the precision in their evaluation, since they typically evaluate the top-ranked terms (up to 300). Recall with respect to an existing terminology resource (term annotations in the PennBioIE and in the Genia corpus) is evaluated in Korkontzelos et al. (2008).

There is no general agreement on the preference of a particular term extraction algorithm. Evaluation results for the same method vary not only with the evaluation metric used, but also with the application domain and the corpus used for extraction.

Termhood methods and the methods measuring the significance of association are found to perform best on a corpus form the European Space Agency using expert judgments on terms (Pazienza, et al., 2005). Hybrid methods, together with termhood methods, performed best on corpora from the life science domain (Korkontzelos, et al., 2008); the PennBioIE corpus (Kulick, et al., 2004), which contains over 700,000 words; and the Genia corpus (Kim, Ohta, Tateisi, & Tsujii, 2003), which contains over 420,000 words.

Zhang et al. (2008) also evaluate term extraction on the Genia corpus and find that tf-idf performs well, but is outperformed by hybrid methods, particularly C-value, which performs best in their evaluation.

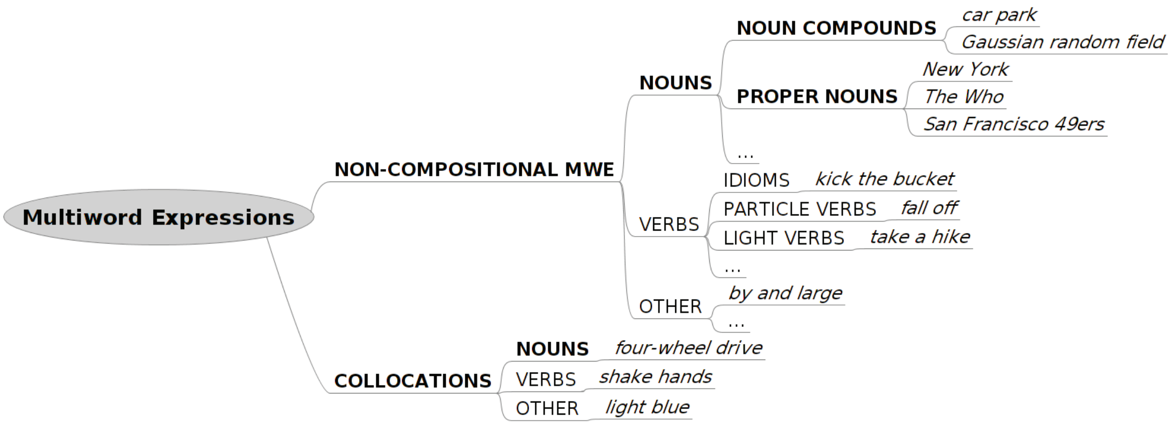
The evaluation by Zhang et al. (2008) explicitly contrasts different types of corpora: term extraction on the Genia corpus is compared to term extraction on a corpus of documents about animals extracted from Wikipedia and consisting of one million words. On the Wikipedia corpus a re-implementation of TermExtractor performs best. The difference in performance on the two corpora is explained with their different composition: C-value performs best on the Genia corpus, which contains a low proportion of unigram terms (reported 11%) and a large number of multiword terms. C-value performs worse on the Wikipedia corpus, which contains a large number of simple terms. Zhang et al. (2008) conclude that the composition of a domain corpus is an important factor in automated term recognition. Unfortunately, they do not present a separate evaluation of the performance on simple terms and multiword terms. Thus it is not clear whether the proposed integrated treatment of simple terms and multiword terms is of advantage.

While there is a lot of evidence in favor of C-value and hybrid methods, these are outperformed by a simple unithood based measure, namely LLR, in an evaluation on PennBioIE (Korkontzelos, et al., 2008).

To sum up, there is no general consensus on a single term extraction method, but there is a tendency to prefer hybrid methods such as C-value. The conclusion to be drawn, however, is that the optimal method is dependent on the particular setting: application domain, type of corpus and term type.

Pazienza et al. (2005) note another aspect of term extraction: they find that besides domain-specific terms, terms from other domains are also detected by the evaluated term extraction methods. This property of term extraction methods is also relevant for the domain independent term extraction setting in this work.

Figure 5. Multiword expression classification



Multiword Expression Mining

Multiword term extraction is closely related to Multiword Expression (MWE) mining in computational corpus linguistics. Therefore, methods from MWE mining have been employed in terminology extraction, like the POS-pattern filtering and the statistical unithood methods introduced above. Multiword expression mining specifically targets unithood, as it aims at the creation of general-language lexicons. For this task, the crucial factor is whether a phrase forms a lexicalized multiword unit in a language; relevance to a particular domain is not required. Thus, a major difference between multiword terminology extraction and multiword expression mining is that general (newswire) corpora are used for mining multiword expressions as opposed to domain-specific corpora in terminology extraction.

Sag et al. (2002) define multiword expressions as "idiosyncratic units that cross word boundaries." Thus, multiword expressions include not only nominal expressions, but also other parts of speech, such as verbs, adjectives, and adverbial phrases. Multiword expressions are interpreted as lexical units, whose irregular semantic, syntactic, pragmatic, or statistical properties justify their own entry in a natural language lexicon. These properties include:

- Semantical non-compositionality: multiword expressions with irregular semantics are *semantically non-compositional*: the meaning of these expressions cannot be inferred from the meaning of their constituent words. Examples are idioms like "to kick the bucket," non-compositional verb-particle constructions like "to give up" or noun compounds like "hot dog."
- Syntactical irregularity: multiword expressions that contain co-ordinations of different parts-of-speech (e.g., a preposition and an adverb in "by and large") are syntactically irregular. Syntactically irregular multiword expressions are typically also semantically non-compositional.
- Statistical irregularity: some multiword expressions are semantically regular, but nevertheless perceived as a linguistic unit, for example "strong tea" or "four-wheel drive." They typically occur together and refer to a particular concept. Consequently, they are considered as institutionalized expressions and are also referred to as *collocations* (Evert & Krenn, 2001).

Figure 5 shows a classification of multiword expressions by part-of-speech and compositionality based on Sag et al. (2002). As multiword terms are multiword expressions, which are relevant to specific domains, all of the listed classes (in capital letters) may contain multiword terms. Some of these will be found only in domain-specific texts (multiword terms in specific sublanguages as in Figure 1), while others also occur in general-language texts (multiword terms placed between general language and specific sublanguages in Figure 1). Classes addressed in the present work in terminology extraction are printed in boldface in Figure 5.

Multiword expressions with different parts-of-speech—potentially divided into subtypes—are usually mined using techniques adapted to the given type/POS (Fazly & Stevenson, 2007). Therefore, various linguistic patterns for candidate identification using POS and syntactic information have been developed. For the identification of multiword expressions—similarly to term extraction—linguistic properties and frequency counts of the candidates and their constituents are taken into account. Additionally, syntactic fixedness and modifiability (Wermter & Hahn, 2005) are features used to distinguish between common natural language phrases and multiword expressions, since multiword expressions have been shown to occur in a narrower range of syntactic constructions and to withstand modifiability: while "kick the bucket" is acceptable, "kick the *big* bucket" is not when employing the idiomatic sense.

Typically, statistical methods measuring the strength of association between a multiword expression and its constituents are used to identify MWEs from corpora: these methods are similar to those for terminology extraction previously discussed. Another group of methods takes context information into account using distributional similarities of multiword expressions: first, context vectors describing the words surrounding the multiword expression candidates are derived from a corpus. Then, these representations are compared to those of their constituent words, using, for instance, the cosine metric, to identify how much the meaning of the multiword expression diverges from the meaning of the constituent words. These measures are used specifically to identify non-compositional multiword expressions such as idioms (Bannard, Baldwin, & Lascarides, 2003; Katz & Giesbrecht, 2006), but often suffer from data sparseness problems.

Another family of multiword expression mining methods exploits translational correspondences between multiword terms and single terms in different languages (Villada Moirón & Tiedemann, 2006): the English "traffic light" is typically translated as one word, "Ampel," in German. Translational correspondences of this kind can be used to identify multiword expressions in different languages using statistical methods. This method is successful in extracting multiword expressions, provided that large parallel corpora are at hand (which can be problematic for certain languages).

The current focus of multiword expression mining is the identification of the best statistical method for particular types of multiword expressions. As with the work in term extraction, different methods are compared: Pecina and Schlesinger (2006) evaluate over 80 statistical methods of MWE extraction on Czech collocations. They find that pointwise mutual information, Pearson's χ^2 test and a version of LLR perform equally well, and almost identically to the best method, which uses distributional semantics.

Additionally, they combine a large number of statistical association measures for multiword expression mining using machine learning techniques. They manage to improve evaluation results significantly from mean average precision of 66% for the best single measures to over 80%.

Within nominal MWEs, the current research is focused on certain types (e.g., noun compound identification (Tratz & Hovy, 2010)) and differentiating between semantically compositional and non-compositional multiword expressions (Korkontzelos & Manandhar, 2009).

Work on single statistical association measures also focuses on improving existing measures: Hoang, Kim, and Kan (2009) include penalization factors for statistical association measures—for instance, to alleviate the bias towards low-frequency terms by PMI; Bouma (2010) tries to avoid inappropriate independence assumptions for statistical association measures by incorporating models of dependence between terms.

Wikipedia as a Knowledge Source for Ontology Construction

In the last few years, Wikipedia, the most successful collaboratively edited encyclopedia, has received wide recognition as a collection of common-sense knowledge and as an information source for various knowledge-intensive technologies. Medelyan, Milne, Legg, and Witten (2009) give an overview of the various uses of Wikipedia and the types of information therein. Two of these are most relevant to this chapter: the taxonomic knowledge and the linguistic knowledge encoded in Wikipedia.

Wikipedia first gained popularity as an alternative to traditional encyclopedias. The quality of content and form has been scrutinized and found to match traditionally edited volumes like the *Encyclopedia Britannica* (Giles, 2005). Wikipedia has the additional advantages of being updated quickly and continuously: the English Wikipedia has reached more than 3.5 million entries in less than 10 years of existence.

Besides English, articles in a large number of languages are provided. They are linked to other languages at the article level. This turns Wikipedia to an interesting resource for multilingual applications and for the projection of language processing techniques from well-resourced to low-resourced languages.

The following information sources relevant for terminology extraction are contained in Wikipedia: *article titles*, the Wikipedia equivalent of encyclopedic headwords, are connected with *article texts*, which contain *definitions* of the titles and detailed descriptions of the article topic. *Links* between pages occur in the article texts. *Disambiguation pages* distinguish between different concepts entered under the same headword. *Redirect pages* introduce variants of an article title—including synonyms and closely related terms, and link to the corresponding article. Through the use of *categories*, articles are also organized in a taxonomy which adds hierarchical structure to the encyclopedia content, and organizes specialized entries under the corresponding, more general entries. Together with the articles and the internal links, the category hierarchy makes up a graph structure, in which concepts are connected by relations. This information can be exploited for relation extraction and ontological structure building.

Also relevant to relation extraction is the information contained in *infoboxes*. These are templates, which introduce attribute–value sets relevant to the topic of the article. Infoboxes are defined for articles belonging to certain categories, for instance locations, animal classes, or natural phenomena. The infobox on the page of a country (for instance, *Italy)* contains a field for the capital (*Rome)* and the currency (*euro*). Thus, semantic relations between concepts are introduced. Infoboxes have been used for tasks such as information extraction and ontology learning.

Wikipedia as a collection of common-sense knowledge backed up by extensive structural information is a good starting point for developing cross-domain and domain-specific ontologies on many subjects. Therefore, it has been exploited for various stages in the ontology construction process, from corpus and terminology extraction to ontology learning.

Corpora extracted from Wikipedia are relevant for terminology extraction, as they contain a large proportion of domain-specific terminology, as well as general-language terms and borderline cases (i.e., terms which occur in domain-specific contexts but also in general language). In evaluations for terminology extraction, they bear comparison with traditional domain-specific corpora (Zhang, et al., 2008; Bonin, Dell'Orletta, Venturi, & Montemagni, 2010). The Wikipedia corpus is either constructed based on a manual selection, using a *Wikiportal* (i.e., a collection of pages on a particular topic area) relevant to the application domain (Bonin, et al., 2010), or using a random selection of articles about animals (Zhang, et al., 2008).

Cui et al. (2008) introduce a more sophisticated approach to the extraction of domain-specific corpora from Wikipedia. Their approach exploits Wikipedia's category labels to extract domain-relevant articles for any given domain automatically. It automatically selects a set of articles relevant to a given root category using only the category information in Wikipedia. First, a so-called classification tree is developed from the root category. It contains the root category, its child categories and articles classified under the categories. The leaves of the tree, Wikipedia articles, are considered as candidates for the domain corpus. They are ranked by relevance to the root category node exploiting linking information between the nodes in the graph.

The next step in the ontology construction process, the extraction of terminology and entities from Wikipedia has been addressed in the context of automatic creation of bilingual dictionaries. Such approaches are usually dependent on parallel corpora, which are often unavailable in specialized domains. Therefore, Wikipedia with its inter-language links and broad coverage of technical domains is a valuable resource for such applications. Exploiting inter-language links appears to be a well-functioning baseline for bilingual terminology extraction, but information from redirect pages and link anchor text has been additionally used to increase the coverage (Erdmann, et al., 2008). Evaluating the extracted resource on a gold-standard dictionary, Erdmann et al. (2008) find that the Wikipedia-based approach compares well to the traditional approach using bilingual corpora, particularly with respect to recall and low-frequency items. Although they also extract multiword terms, Erdmann et al. (2008) evaluate only single words and do not consider multiword terms in their evaluation. They expect even better improvements using Wikipedia for the extraction of multiword terms compared to standard techniques and their Wikipedia baseline, but cannot prove this assumption.

Wikipedia has also been used for a task related to terminology extraction—namely recognition of named entities, which covers person names, location names and the like. Named Entities (NEs) are relevant to ontology construction, since they represent instances of ontological concepts. They are sometimes covered by terminology extraction, but, compared to ordinary terms, require special treatment: besides identifying word sequences as named entities, classification into NE types and disambiguation of NEs are required, as in the case of the person name "George Bush," which could refer to either the 41st or the 43rd president of the United States. The approach proposed by Cucerzan (2007), for instance, employs Wikipedia for the identification and disambiguation of named entities. First, a dictionary of named entities is created by collecting article titles and their spelling variants from redirects and link anchor texts. Disambiguation and classification information is then extracted from redirect pages, disambiguation pages, category tags, and using "list of *" entries in Wikipedia articles, where * represents a named entity category or a subtype (e.g., "list of countries"). Additionally, contexts of the extracted NEs are stored. Using this information, spelling variants of a NE are associated with a particular entity, and classification of this entity is performed. A new occurrence of a named entity can then be disambiguated by comparing its context with the Wikipedia article text of the candidate entities and the context information stored in the dictionary.

Wikipedia has also been subject to various ontology learning efforts, for instance the YAGO (Suchanek, Kasneci, & Weikum, 2007) and DBpedia (Auer, et al., 2007) projects. They aim at alleviating the coverage bottleneck of expert-built, handmade ontologies, like CYC (Lenat, 1995), and taxonomic resources, like WordNet (Fellbaum, 1998).

The examples introduced in this section show that Wikipedia contains a wealth of information relevant to ontology learning. Information in Wikipedia can be mined from article texts, infoboxes, and from structural elements, such as the internal link and category structures. All of these elements have been exploited for the semi-automatic construction of ontologies, either in the automatic creation of taxonomies and other structured resources, in the generation of terminological dictionaries and named entity gazetteers, or in the creation of domain-specific corpora. In the next section, we will present another application of Wikipedia as a knowledge source for ontology construction by introducing our work on extracting multiword terminology from Wikipedia.

MINING MULTIWORD TERMS FROM WIKIPEDIA

Motivation: Wikipedia as a Source of Multiword Terms

Wikipedia has been shown to be a valuable resource in ontology construction. In this work, we particularly focus on those properties of Wikipedia relevant for the extraction of multiword terms.

We assume that multiword terminology extraction needs to be treated differently from the extraction of unigram terms. This is backed up by previous work on term extraction techniques, such as the work by Zhang et al. (2008), who suggest an integrated approach for unigram and multiword terms, but find that some techniques work better than others, depending on the proportion of unigram and multiword terms in the source corpora: they report that C-value—better equipped to deal with multiword terms, as it takes nestedness of terms into account – performs better on the Genia corpus (which contains a large proportion of multiword terms), than tf-idf, a measure that does not treat multiword terms different from unigram terms. Tf-idf in contrast performs better than C-value on a corpus with a large proportion of unigram terms. We conclude that optimal results could be achieved by extracting multiword terms and unigram terms separately, using appropriate methods for both.

We use Wikipedia as a knowledge source for the extraction of multiword terms for two reasons: first, it supplies human-generated markup which can be exploited for candidate extraction, and second, it is a valuable resource of domain-specific terminology and general world knowledge. With our approach, we extract domain-specific terms, but also multiword expressions found in general language. This is motivated by the fact that, as shown in Figure 1, the decision on the domain relevance of a term is not clear-cut. Wikipedia is expected to contain highly domain-specific multiword terminology, less specialized multiword terms relevant to various domains, and also general-language multiword expressions. Conventional methods of term extraction from domain-specific corpora often aim at excluding the third class: the tf-idf measure, for instance, penalizes terms, which occur in many documents in the corpus and are therefore considered less domain-specific; Bonin et al. (2010) use general-language corpora specifically to filter out general-language terms. There are, however, application scenarios, in which terms closer to general language cannot be neglected. One example is the creation of a medical knowledge base to be queried by lay persons: both specialized technical terminology and colloquial expressions referring to diseases or bodily functions are of relevance in such an application. Another application scenario which requires a term vocabulary covering various degrees of expertise and specificity is in the e-learning domain, where knowledge on various topics is presented to students of varying degrees of expertise. Moreover, the domain boundaries are more blurred for applications in e-learning than for many traditional applications of ontologies. Therefore, resources transcending traditional domains or study subjects are required. Being able to specify relevant domains on the fly, using only a seed list of domain terms or a domain corpus as input, is an additional asset of Wikipedia as an information source.

Summing up, domain relevance of a term greatly depends on the target domain and application. We therefore present a high-recall approach to extract a large domain-transcending resource of terms of varying domain-specificity from Wikipedia together with additional information, such as categories and definitions of terms that can be used to filter the terms with respect to specific domains and application scenarios. In the following sections we present the construction of the resource: term candidate identification, ranking and term selection, extraction of additional information and the evaluation of the extracted terms.

Candidate Extraction

We target the problem of phrase boundary identification for multiword term extraction by tapping into human knowledge encoded in Wikipedia markup: we rely on phrase boundaries explicitly marked by humans. These are word sequences marked by different typesetting (bold, italics), or wiki markup (link anchor texts, titles, headers). This is similar to work in automated term extraction from web texts which uses XHTML markup to identify phrase boundaries (Brunzel, 2008), and to work in bilingual terminology extraction which exploits Wikipedia's inter-language links to extract bilingual term pairs (Erdmann, et al., 2008; Erdmann, et al., 2009).

We extracted multiword term candidates from two data sources within the English Wikipedia using a Wikipedia dump from 2007 and the Java Wikipedia API (Zesch, Müller, & Gurevych, 2008) as a toolkit. The first data source is the set of Wikipedia *article titles*; the second source is the text of Wikipedia articles. We used article titles directly as term candidates, without further processing. From the article text, multiword term candidates were extracted using the following set of MediaWiki markup patterns,

- Anchor Text (Internal Links): [[target|**term_candidate**]]
- Section Headers: ===* **term_candidate** ===*
- Phrases in Boldface: '''**term_candidate**'''
- Phrases in Italics: "**term_candidate**"

whereby **term_candidate** is defined as the sequence of two or more *words*, (sequences of characters, including numerals, hyphens, and apostrophes) separated by spaces.

Figure 6 lists the number of extracted term candidates by term size (i.e., the number of constituent words in a term) for Wikipedia titles. More than 40% of the over 3.3 million titles consist of two words, compared to 17.5% unigram titles. The multiword titles constitute 82.4% of the total, and those consisting of two to four words still represent 72.7% of all titles. Only 10% of the titles consist of terms longer than four words.

We restricted the size of term candidates extracted from the Wikipedia articles to two to four constituent words. We had several reasons for this filtering by term size: first, as the Wikipedia titles show, candidates consisting of two to four words were the majority of the extracted term candidates. Second, longer phrases, which were likely to occur in a larger proportion among the term candidates extracted from Wikipedia articles (marked by link anchor text, headers and special typesetting), contained full sentences or citations, which we did not target in our experiments. To ease the effort involved in further processing, we excluded these. A third reason for the size filtering

Figure 6. Term candidate statistics (Wikipedia titles)

# constituents	# terms	% of total
1	582,469	17.5
2	1,354,349	40.9
3	719,062	21.7
4	334,531	10.0
5	161,012	4.9
6	83,429	2.5
7	38,406	1.2
...	...	...
total	3,312,743	100.0
2+	2,730,274	82.5
5+	322,332	9.7
2–4	2,407,942	72.7

is that we aimed to alleviate the effects of term size on the statistical ranking.

The following filter was applied to all the extracted word sequences: multiword term candidates were not allowed to contain punctuation marks except for the following signs: '`&%@-. Additionally, they were required to start with an alphanumeric character. We applied case folding, i.e., all candidates were lowercased, to avoid additional efforts of case normalization. This strategy made subsequent processing, such as the collection of term frequencies, easier.

Thus, we extracted more than 5 million multiword term candidates of size two to four. Of these, 1.6 million stem from titles and 4.3 million from markup in Wikipedia articles. The lower number of candidates from titles compared to the raw numbers in Figure 6 is due to the applied filters and lowercasing.

Note that this step did not require any linguistic information besides heuristics on term composition and word separation in English. Thus, our approach of term candidate extraction could easily be applied to other languages in Wikipedia.

Candidate Ranking

The quality of a term extraction process which relies only on Wikipedia-based filtering is quite high already—manual inspection of the extracted term candidates revealed a large number of domain-specific and general-language terms. We nevertheless apply a ranking step to filter out ungrammatical sequences ("amount of prize") and regular English phrases ("married couples"), because we expect them to receive a low score in the ranking.

Therefore, we combine our technique with statistical methods typically used for the extraction of multiword terms. Since we do not specifically focus on term extraction in a particular domain, but also include terms closer to general language, we apply a statistical association measure proven to be efficient for the extraction of multiword expressions from corpora, namely pointwise mutual information (Hoang, et al., 2009).

Pointwise Mutual Information (PMI) measures the strength of association between the constituent words of a multiword term candidate in a corpus by comparing the expected probabilities of the multiword term to the probabilities observed in the corpus. Expected probabilities are computed as products of the probabilities of the constituent words, assuming independence between the constituent words (see Figure 6). PMI is interpreted as follows: a high PMI value shows a strong association between the constituents of the candidate terms, and thus provides evidence, that they indeed constitute a multiword term.

The PMI measure is usually applied to bigram candidates. It needs to be adapted to appropriately deal with terms of longer size. Therefore, several options have been suggested (Da Silva & Lopes, 1999; Korkontzelos, et al., 2008): the standard application of the PMI measure compares the observed probabilities to the expected probabilities modeled as the product of the probabilities of the two constituent words w_1 and w_2 of a term:

$$\mathrm{PMI}(w_1 w_2) = \log \frac{P(w_1 w_2)}{P(w_1)P(w_2)}$$

For term candidates consisting of three or more words ($c = w_1,\ldots,w_n$), there are several options to compute the expected probabilities of the multiword term. The easiest one is applying the approach for bigrams and calculating the expected probabilities as the product of the observed constituent probabilities under the assumption that the constituents of the n-gram are independent of each other:

$$\mathrm{PMI}_{naive}(w_1,\ldots,w_n) = \log \frac{P(w_1,\ldots,w_n)}{\prod_{i=1}^{n} P(w_i)}$$

This assumption is criticized as being inadequate for bigrams (Bouma, 2010) and even more problematic for longer terms, since it does not take the phrase structure of multiword terms into account: 3-gram multiword terms are usually made up of a single word and a bigram ([Gaussian [random field]]), 4-gram multiword terms of two bigrams ([[finite dimensional] [vector space]]) or a 3-gram and an unigram ([[raster to vector] conversion]).

Da Silva and Lopes (1999) suggest a way of computing the expected probabilities for longer term candidates. It is called "fair dispersion normalization" and involves splitting longer n-grams into "pseudo-bigrams" using all possible split points and using the average of the probabilities as expected probabilities for the n-gram:

$$\mathrm{PMI}_{naive}(w_1,\ldots,w_n) = \log \frac{P(w_1,\ldots,w_n)}{\prod_{i=1}^{n} P(w_i)}$$

For the multiword term candidate "Gaussian random field," the fair dispersion normalization would compute the average of the observed probabilities for the split [[Gaussian random] field] and [Gaussian [random field]].

A simpler variant, called "pessimistic split," uses the split with the highest observed likelihood (Korkontzelos, et al., 2008), in our example [Gaussian [random field]]:

$$\mathrm{PMI}_{pess}(w_1,\ldots,w_n) = \log \frac{P(w_1,\ldots,w_n)}{P(w_1,\ldots,w_i)P(w_{i+1},\ldots,w_n)}$$

Here, the split point i is determined as the one maximizing $P(w_1,\ldots,w_i)P(w_{i+1},\ldots,w_n)$. For this strategy, a comparatively high number of occurrences is required to receive a high PMI, so it leads to a conservative decision: if a candidate receives a high ranking using pessimistic split, it is very likely that the candidate actually is a collocation. In our evaluation, we compared both of those normalization strategies.

The collocation measure relies on corpus frequencies of the multiword expression candidates. Two benefits of using term candidates from Wikipedia are the good coverage of technical domains and neologisms. We use the Wikipedia text as a corpus for the candidate ranking, since we do not expect to find similar coverage on technical terms and neologisms in the newspaper corpora typically used for this task. Therefore, we extracted the counts for all extracted term candidates from Wikipedia texts. To collect the counts, we considered only the cleaned text without wiki markup. Additionally, we extracted counts for the subsequences of terms with more than two constituents. These were required to compute the normalized PMI scores for term candidates of size three and four.

We restrict minimum occurrences to accommodate the bias of PMI to prefer lower frequency items, as suggested by Pecina and Schlesinger (2006): only those candidates with at least six occurrences in the Wikipedia corpus were considered for ranking. Using a corpus as large as

Wikipedia, the slightly lower recall resulting from the frequency filtering is not an issue for us. Out of a total of 5.26 million multiword term candidates, a ranking was computed for 1,032,859. The size reduction is mainly due to the frequency cutoff. Besides, a few subsequences of terms were not found in the Wikipedia corpus because of errors in the automatic removal of wiki markup.

We found that term candidates with more than two constituents (i.e., 3-grams and 4-grams) receive both high and low positions in the ranking. This observation indicates that the applied normalization of the PMI measure works well: these longer terms are neither collectively favored nor disfavored. The fair dispersion and pessimistic split normalization provide very similar results (Spearman's rank order correlation between these measures being 0.996); therefore we proceeded with analysis and further processing based on the latter method.

Manual analysis of the ranking showed that the top ranks are mainly given to named entities, such as names from the scientific classification of plants and animals ("archaeocydippida hunsrueckiana," "suricata suricatta"). The lowest scores were given to ungrammatical phrases and misspellings ("would of"), or names and phrases that appear as such ("the who"). The middle ranks were occupied by multiword terms of varying compositionality ("swell box," "utility pole," "dog whistle," "aramaic speaker"), named entities ("cable guy," "milford railway station"), and specialized terminology ("sister clade").

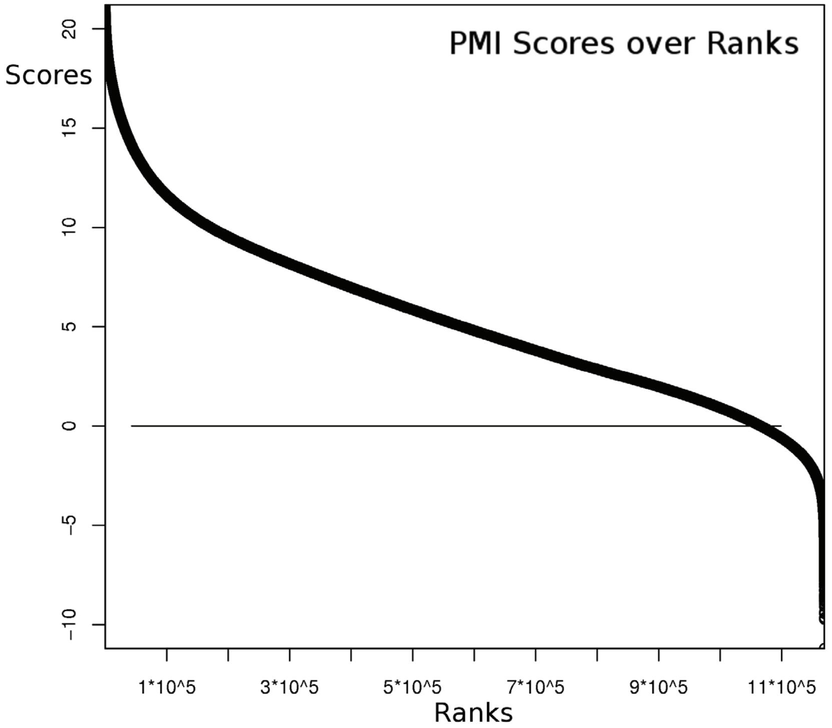
To decide which candidates to admit to the final resource, we determined a cutoff value. Observed PMI scores range from −7.74 to 18.43. Since multiword term candidates have already been selected by human Wikipedia authors, either by marking up the candidate or specifying it as an article title, comparatively high quality can be expected in the resource. Therefore, we need only remove candidates with really low PMI scores from the full set. A score-over-rank plot of multiword term candidates (see Figure 7) suggests a cutoff at PMI=0. Therefore, all candidates with a score higher than 0 are selected for the multiword term resource. About 29,000 term candidates are thereby discarded; 1,003,508 remain.

Candidate Selection and Filtering

The set of selected candidates contains many different types of multiword terms: there is a large number of named entities as well as technical terms and general-language terms of varying compositionality (including non-compositional multiword expressions and collocations, see Figure 5). We performed automatic and manual analysis to classify the selected multiword terms and to get an estimate of the distribution of multiword term types in the resource.

During the extraction of term candidates, we tagged the Wikipedia text corpus with POS tags. A multiword term candidate was associated with the most frequently occurring POS sequence. The Stanford named entity tagger (Finkel, Grenager, & Manning, 2005) was used to assign general named entity tags (Person, Location and Organization) to occurrences of the term candidates. We use this information to divide the set of multiword terms into named entities and other terms. First, all multiword terms which have a corpus occurrence tagged as a named entity sequence are classified as named entities. Additionally, a particular sequence of POS tags was used to identify named entities missed by the Stanford NE tagger, such as film titles. The "proper noun" tag in the tag-set used refers to named entities and manual analysis showed that terms tagged as proper nouns are likely to be named entities. Therefore, terms which were tagged as a sequence of at least two proper nouns (NP, NPS), optionally modified by determiners (DT), adjectives (IN) and conjunctions (CC) and ending on a proper noun, identified with the pattern "(NP|NPS) ((CC|DT|IN|NP|NPS))*(NP|NPS)," were also classified as named entities. These are more than the half of the multiword term candidates surviving the PMI cutoff. We used linguistic

Figure 7. Score over rank plot (PMI ranking)



processing tools for POS tagging and named entity identification in our work. These could, however be replaced by language-independent approaches for named entity identification using structural information from Wikipedia, for instance using the technique suggested by Richman and Schone (2008) for multilingual named entity recognition in Wikipedia. Thus, our approach could easily be applied to languages other than English, for which language-dependent POS taggers and named entity recognizers are not available.

Multiword terms not classified as named entities were subject to additional filtering steps: they were filtered based on a set of heuristics in order to exclude what we call "Wikipediaisms"—expressions typical for Wikipedia which therefore receive a high score in the ranking. Examples include the phrase "external links" and multiword units of the form "lists of X" (e.g., "lists of countries"). Additional filtering based on POS sequences was performed to exclude ungrammatical phrases, such as those starting with conjunctions or ending with definite or indefinite articles. Unlike previous work, we did not use a positive list of POS patterns for the extraction of multiword term candidates as this would exclude a wide range of multiword terms.

Properties of the Resource

The resource mined from Wikipedia contains more than 880,000 terms and consists of two parts: one part containing 528,536 named entities, and a second part containing 356,467 Multiword Terms (MWTs). We refer to the former as NE resource and the latter as MWT resource. Both resources are available for download in XML format at http://www.ukp.tu-darmstadt.de/data/multiwords. Properties of the resources, as well as details on the information they contain will be presented in this section.

The terms in the two resources show a different distribution regarding their source in Wikipedia. The Named entities are mainly found in Wikipedia titles: 76% of the entries in the NE resource originate from Wikipedia titles, 24% from markup, while (only) 45% percent of the MWT resource were mined from Wikipedia titles, and 55% percent come from various markups. This shows that Wikipedia markup is an important source of multiword terms and should not be neglected in favor of titles. Out of the markup terms in the MWT resource, roughly 55% occur as link anchor text and 45% as highlighted text or headers. Both parts of the resource show a similar distribution of term size. 73% of the terms in the full resource consist of two words, 21% of three words and 7% of four words.

Besides the information on the source of the term in Wikipedia, the resources provide POS and frequency information, and the PMI score.

In order to ease the integration of the Wikipedia-based multiword term resource into semi-automatic ontology construction, we augmented our resource with information from Wikipedia: we extracted definitions and category tags for those multiword terms in the MWT resource that can be associated with Wikipedia articles. Definition information and category information can be used to integrate terms into an existing ontology based on semantic similarity and to ease the establishment of relations between terms for a new ontology. Furthermore, the provided information can be used for adaptation of our domain-transcending resource to particular target domains—e.g., through the filtering for predefined categories.

Category tags added by Wikipedia editors are provided for each article and can be directly extracted for each article. The article on "gross domestic product" is tagged with the following categories: "*Index numbers"* and *"National accounts."* Based on these categories, the term can be classified as belonging to the economy and finance domains.

While category tags are easy to obtain, extracting definitions from Wikipedia requires some understanding of the structure of Wikipedia articles. We interpret the first paragraph of a Wikipedia article as definition of the associated concept (Zesch, Gurevych, & Mühlhäuser, 2007), as it is typically used to introduce the article title. An example is the following text section for the Wikipedia article on "gross domestic product":

The Gross Domestic Product (GDP) or Gross Domestic Income (GDI), a basic measure of a country's economic performance, is the market value of all final goods and services made within the borders of a nation in a year. GDP can be defined in three ways, all of which are conceptually identical. First, it is equal to the total expenditures for all final goods and services produced within the country in a stipulated period of time (usually a 365-day year). Second, it is equal to the sum of the value added at every stage of production (the intermediate stages) by all the industries within a country, plus taxes less subsidies on products, in the period. Third, it is equal to the sum of the income generated by production in the country in the period - that is, compensation of employees, taxes on production and imports less subsidies, and gross operating surplus (or profits).

For some Wikipedia articles, the first section does not provide a textual definition, or only a very short text (less than 100 characters). In this case, we added the next section to the definition. In the resource, we highlight, when more than the first section was used to compile the definition.

Collecting category and definition information is straightforward for the 161,072 terms originating from Wikipedia titles. Additionally, the 195,395 terms from link anchor text can be associated with the target of the link. This allows us to associate

Box 1.

```
<mwe>
        <lemma> gross domestic product </lemma>
        <pos> JJ JJ NN </pos>
        <freq> 613 </freq>
        <pmi> 9.958813181942412 </pmi>
        <source> wiki_titles </source>
        <sense>
                <page_title> Gross domestic product </page_title>
                <category> Index numbers </category>
                <category> National accounts </category>
                <category> Gross Domestic Product</category>
                <category> All articles with unsourced statements </category>
                <definition_first_paragraph> The Gross Domestic Product (GDP)
or Gross Domestic Income (GDI), a basic measure of a country's economic per-
formance, is the market value of all final goods and services made within the
borders of a nation in a year.  GDP can be defined in three ways, all of which
are conceptually identical. First, it is equal to the total expenditures for
all final goods and services produced within the country in a stipulated pe-
riod of time (usually a 365-day year). Second, …
                </definition_first_paragraph>
        </sense>
</mwe>
```

link anchor texts with definitions and categories of their link targets, unless they link to overview pages such as "lists of X." Often, link anchors are associated with several targets, leading to different Wikipedia pages. These anchors can be considered ambiguous and receive several senses in our resource. Wikipedia titles can also be ambiguous, if they point to a disambiguation page.

The example entry for "gross domestic product" shows the information that our resource provides on the term in XML format. The term is associated with just one sense in Wikipedia, for which category information and a definition extracted from the first paragraph are available (see Box 1).

Summing up, we managed to enrich our MWT resource with over 240,000 definitions for multiword terms in the resource. We also extracted over 460,000 category tags. 148,793 of the multiword terms in the resource are tagged with an average 3 categories and provide a definition text.

Using the Resource for Ontology Construction

The enriched resource provides additional information which is typically not available when taking the standard approach of extracting terminology automatically from domain corpora. This information can be exploited in various tasks in the ontology construction process like 1) filtering (domain-specific terms can be extracted based on category information), 2) adding textual descriptions (definitions can serve as a starting point to provide concise descriptions in the ontology) or taxonomy construction (Wikipedia categories are useful for the hierarchical structuring of the terms). Furthermore, our resource is domain transcending, which we consider a great benefit for the development of new ontologies: instead of repeating the terminology extraction process for every new domain, the existing resource can be adapted to a particular domain using a domain filtering approach, which is less time and resource demanding.

Evaluation: Annotation Study and Comparison

In order to get a better estimate of the different types of multiword terms which constitute our resource, we performed an annotation study. The study was performed specifically to evaluate the MWT resource, as we are particularly interested in the quality of our resource for terminology extraction rather than named entities. We focused on the assessment of the unithood, rather than the termhood, of the extracted multiword terms. The reason for this is twofold: first, a high proportion of domain-relevant terms is to be expected in our resource due to Wikipedia's encyclopedic nature; second, as the resource is domain-independent – it is intended to cover a broad range of domains and terms of varying degrees of specialization – relevance with respect to a particular domain is not a good evaluation criterion in our case. Therefore, we relied on a classification frequently used for the evaluation of general-language multiword expressions. It classifies terms as being either

1. *non-compositional*, which covers phrases whose meaning cannot be completely inferred from the meaning of their parts and typically includes technical terminology,
2. *collocations*, which can be understood based on the composition of the constituent terms, but is lexicalized (and thus a useful candidate for ontology construction),
3. *regular phrases*, which is not considered lexicalized, or
4. *ungrammatical*.

Out of the 356,467 multiword terms in the MWT resource, we sampled 2500 randomly and had two human annotators annotate each term with one of these four classes. The annotators were additionally asked to mark terms which they considered named entities. This was done to identify named entities that slipped through the filtering step.

The annotators were equipped with a detailed annotation guide containing examples of the different classes and criteria and tests for identification. Furthermore, they were asked to perform a quick web search for each expression in order to get familiar with unknown terms and discover named entity usages of terms which would otherwise have been classified as regular phrases or as ungrammatical. (This frequently applies to film titles.)

To estimate the quality of the annotation, we measured the agreement between the annotators. For the binary NE classification ("NE" vs. "not NE"), we computed simple agreement: 0.87 of all rated terms receive the same rating in the NE-dimension. Terms identified unanimously as NE by the raters were not taken into account in the further agreement evaluation.

For the three termhood classes, we computed the κ-score value between the two annotators as suggested by Krippendorf (1980). The agreement score between the two annotators is κ=0.42 on the three-class rating ([1] vs [2] vs [3]). When considering only the binary classification into regular phrases [1] and valuable multiword terms (classes [2] and [3] together), agreement is κ=0.48.

This value is considered fair agreement according to the scale used by Landis and Koch (1977) and low agreement according to Krippendorf (1980). These results have to be interpreted in light of the difficulty of the rating task: the boundaries between the classes are often not clear cut and the raters cannot be expected to be familiar with all domains in Wikipedia. This means that low annotator agreement does not imply a low quality of the resource. Determining whether an expression is compositional is a difficult task, which is even more difficult in Wikipedia due to the large number of domain-specific terms. The fact that terms from Wikipedia are difficult to annotate suggests that they contain a high variety of technical terminology, which is desired for the construction of a terminological resource. Differing background knowledge of the annotators may lead to different annotations—for instance, when an annotator is familiar with a certain domain she may annotate a phrase as compositional, because the sense of the headword of the term is familiar, while the other annotator annotates the same phrase as non-compositional based on the more prominent prototypical meaning of the head word. One example is "gross national product" which can be understood as compositional by someone familiar with the financial domain and meanings of "gross" and "national product" in this domain. A person only familiar with more colloquial senses of "gross" (e.g., being bulky or disgusting) is more likely to classify "gross national product" as non-compositional. Another example is "plain dress," which can be understood compositionally as referring to an unadorned dress, while a person familiar with the lifestyle of the Amish will understand it to refer to their sumptuary rules, including particular modesty and conservative cut. The identification of named entities also relies on the familiarity of the annotators with geographical locations and company names, or the results that ranked top in the web search performed by the annotator.

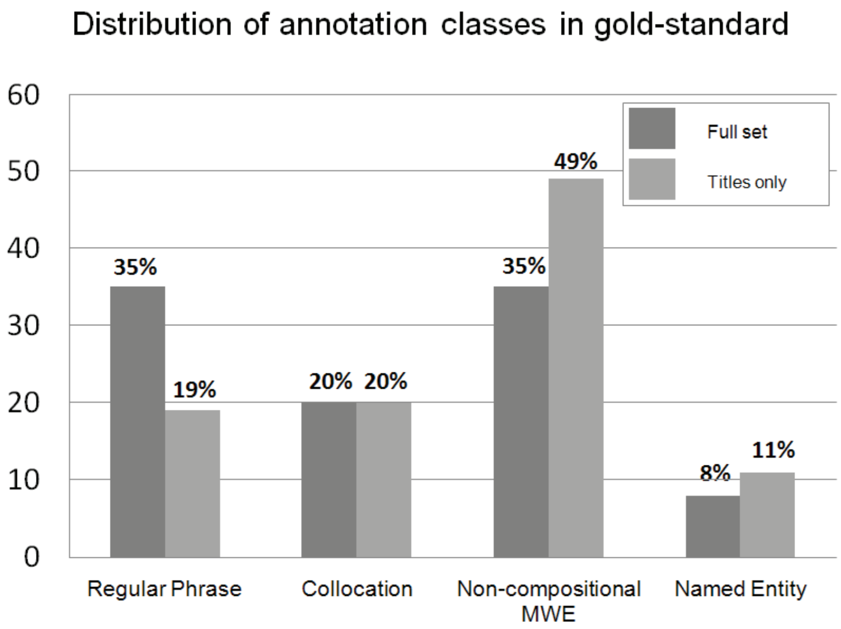
In order to create a reliable annotation on the full set, the multiword terms on which the first two annotators disagreed were re-evaluated by the expert annotator who also designed the annotation guide. The third annotator agrees well with the two initial annotators, where the latter agree: the κ-score between the third rater and the first two raters is 0.69 on 3-class evaluation and 0.74 on binary evaluation. The score was computed based on a random sample of 150 terms from the set of terms on which the first two annotators agree. The fair agreement justifies using the expert rater's disambiguation of the terms on which the first two raters disagreed.

Terms which were rated as NE by both annotators were considered to be reliable NEs and therefore not re-evaluated by the expert rater. The unanimous initial ratings and the corrected annotations by the expert were used to compile a gold-standard annotation.

All together, 891 terms of the sample are classified as non-compositional, 505 as collocation, 881 as regular phrase, and 220 as NE. Three terms were identified as ungrammatical. The evaluation shows that the multiword term resource contains some noise, but still a large amount of valid terms. Figure 8 presents the percentage distribution on term classes in the evaluated set. Besides 8% named entities, which are also useful for ontology construction (albeit not the focus of our evaluation), the gold-standard set contains more than 55% valuable multiword terms (i.e., non-compositional multiword expressions and collocations). The same proportion can be expected in the full MWT resource. Moreover, terms which are neither classified as non-compositional or collocations, nor identified as a named entity cannot generally be considered noise from the point of view of offering a "multiword term resource," as manual evaluation on a sample of 200 terms showed that 75% of the regular phrases nevertheless represent domain concepts (e.g., "wheat field" in an agriculture setting). Thus, more than 81% (around 290,000) of the 356,467 multiword terms in our resource can be expected to be valuable terms.

Figure 8 compares the distribution of term types in the evaluated sample. The configuration of multiword term sources we used, Wikipedia markup and titles, is compared to the baseline of using titles only. Analyzing the gold-standard with respect to the source of the terms shows that the proportion of non-compositional multiword terms is higher when considering only the multiword terms stemming from Wikipedia titles: as shown

Figure 8. Distribution of term types in evaluated sample



in Figure 8, 49% of the 1146 title terms are non-compositional, but only 19% of them are regular phrases. Thus, the subset of terms in the resource stemming from titles, consisting of over 160,000 multiword terms, could be used for applications which require higher precision than recall. Splitting the markup terms further, into those originating from anchor text and other highlighting (headers and typesetting), would allow for a finer grained tuning of precision versus recall, as terms from anchor texts contain only 39% regular phrases, compared to over 50% for terms from headers and typesetting.

Focusing on precision however means a loss in recall: in the gold standard, 54% of the terms stem from titles and 46% of the terms from markup. In the whole MWT resource, multiword terms originating from Wikipedia markup, text highlighting and anchor text constitute more than the half, 55% of the resource. Not considering these would reduce the size and coverage of the resource dramatically.

To further evaluate the Wikipedia-based resource, a comparison was performed with the multiword expressions contained in the Princeton WordNet (Fellbaum, 1998), an expert-built lexical semantic resource. WordNet contains a taxonomy of lexical entries. The basic structure is the *synset*, grouping all lexical entries considered synonymous to each other. Synsets are equipped with short definition glosses and inflectional information. Synsets are connected to each other via relations like hyponomy ("is-a") and antonymy, thus spanning a graph structure. WordNet has been widely used to compute similarities between words and to perform word sense disambiguation.

The comparison shows first of all that the multiword resource extracted from Wikipedia exceeds WordNet dramatically with respect to size, and second that a large proportion of the expressions in WordNet is covered by the Wikipedia-based resource. WordNet contains over 110,000 lexical entries on nouns, 68,000 of which are nominal multiword expressions. Of these, about 63,000 have the same size as the multiword terms in the Wikipedia-based resource—namely, two to four constituent terms. The Wikipedia-based resource covers over 70% of these 63,000 multiword expressions. This gives additional proof that a large number of high-quality terms are contained in the Wikipedia-based resource.

FUTURE RESEARCH DIRECTIONS

In this work, we specifically focused on the extraction of a multiword term resource and exploited properties of Wikipedia that are beneficial to this particular task. In future work, we plan to extract unigram terms as well, which would be a valuable extension to our Wikipedia-based resource of multiword terms. For the ranking of unigram terms extracted from Wikipedia, we will employ methods appropriate for unigram terms, focusing on termhood rather than unithood.

Another natural direction of future research is the implementation and evaluation of the domain filtering: we plan to evaluate our Wikipedia-based resource in the e-learning domain. Therefore, we will develop techniques to optimally select terms relevant to a particular domain (i.e., *domain filtering*) using category information and definitions from Wikipedia. We consider several options for domain filtering: the first is to select a seed set of categories relevant to the target domain and use it to extract relevant terms from the resource. Additionally, the category hierarchy can be exploited to expand the set of seed categories by their child categories, similar to the work by Cui et al. (2008).

The second option is to use a seed set of domain-relevant expressions as a domain filter: semantic similarity between the seed expressions and the information extracted from Wikipedia, especially definitions, but also categories, can be used to identify terms as relevant to the domain. A collection of domain-specific texts can be alternatively used to filter the resource: terms from the resource occurring in the texts are likely to be domain relevant. This strategy requires the least effort, if a collection of relevant texts is available.

The implementation of the filtering method will allow us to further evaluate the resource quality with respect to particular application domains. In this context, an important aspect of terminology extraction for ontology construction is not only the domain relevance of the extracted terms, but also whether the domain is represented well by the terms in a resource (Zhang, Xia, Greenwood, & Iria, 2009). The former can be evaluated using precision, while the latter is related to recall: not only should the resource cover a large number of relevant terms, but *all relevant aspects* of the domain should be represented in the terminology resource. To evaluate this, a domain-specific text corpus can be applied: if there are documents in the corpus, which are not covered well by the resource—i.e., which do not contain any resource terms—or only terms generally relevant to the domain (as shown by high document frequency, the number of documents in a collection a term occurs in), the domain might not be covered adequately by the resource. Adequate domain coverage, however, is an essential feature of terminological resources, because omissions introduced in this foundational step of semi-automatic ontology construction are propagated throughout the construction process and affect the final quality of the ontology.

We furthermore plan 1) to evaluate our method for extracting multiword terms from Wikipedia on languages other than English, to prove the language independence of our approach, and 2) to evaluate statistical ranking methods besides PMI. Finally, we plan to make the software packages for language independent extraction of multiword terms from Wikipedia available for research purposes.

CONCLUSION

The efficient creation of high-quality terminological resources is a foundational step in semi-automatic ontology construction. In this chapter, we presented a method using the collaborative encyclopedia Wikipedia, which has received wide recognition as an information source of domain-specific and general world knowledge, for the creation of a domain-transcending multiword term resource. Wikipedia offers several advantages as a resource for term extraction: 1) high coverage of specialized domains, 2) quick evolution with respect to emerging research areas, 3) domain information in the form of a category hierarchy, and 4) its multilinguality and intra- as well as interlingual link structure. Moreover, Wikipedia offers structural elements (markup) supporting the task of phrase boundary detection, which is essential to multiword term extraction. Exploiting the markup for the extraction of term candidates, our method is basically language-independent, as it does not require language-specific processing tools. These tools can, however, be employed to further enhance the precision of the extraction process, for instance by applying POS-filters to term candidates and thus considering only grammatical phrases of a particular language.

Our method provides a set of over 500,000 NEs, including general categories like persons, locations, and NE types relevant to specific domains (film titles, chemical substances, etc.). Beyond that, we managed to extract over 350,000 multiword terms using our approach. For these, we additionally extracted category and definition information associated with the terms from Wikipedia and provide it as additional information as part of the resource. To evaluate the multiword part of the resource, a sample of terms was annotated for four linguistic classes of unithood by human raters. Focusing on unithood rather than termhood is motivated by the need of a domain independent evaluation of our domain transcending resource. Moreover, the extraction of a large proportion of domain-relevant terms is inherent to our method due to using Wikipedia as an information source. According to the evaluation, 55% of the terms in the sample are non-compositional multiword expressions and collocations, 35% regular phrases and 8% named entities (that could belong to the NE part of the resource, but were missed by our filters). While *regular phrases* are not considered lexicalized units in the annotation setting (i.e., they are neither semantically non-compositional nor institutionalized) manual analysis showed that over 75% of them nevertheless represent domain concepts (e.g., “wheat field” in an agriculture setting). This means that while the multiword term resource contains some noise it still holds a large amount (some 89%) of valid terms. The coverage of the resource also compares favorably to other dictionaries of multiword terms like WordNet. The entire resource, with all the information described here, is freely available at http://www.ukp.tudarmstadt.de/data/multiwords.

ACKNOWLEDGMENT

This work has been supported by the Emmy Noether Program of the German Research Foundation (DFG) under the grant No. GU 798/3-1 and by the Volkswagen Foundation as part of the Lichtenberg Professorship program under grant No. I/82806. We would like to thank the anonymous reviewers for their valuable comments and our colleagues at the Ubiquitous Knowledge Processing Lab for many fruitful discussions during the work on this chapter.

REFERENCES

Ahmad, K., Gillam, L., & Tostevin, L. (1999). University of Surrey participation in TREC-8: Weirdness indexing for logical document extrapolation and retrieval (WILDER). In *Proceedings of NIST Special Publication 500-246 the Eighth Text REtrieval Conference (TREC 8)*, (pp. 717–724). Retrieved from http://trec.nist.gov/pubs/trec8/papers/surrey2.pdf.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In Aberer, K. (Eds.), *The Semantic Web* (*Vol. 4825*, pp. 722–735). Lecture Notes in Computer Science Berlin, Germany: Springer. doi:10.1007/978-3-540-76298-0_52

Aussenac-Gilles, N., Biébow, B., & Szulman, S. (2000). Revisiting ontology design: A method based on corpus analysis. In Dieng, R., & Corby, O. (Eds.), *Knowledge Engineering and Knowledge Management Methods, Models, and Tools* (*Vol. 1937*, pp. 27–66). Lecture Notes in Computer Science Berlin, Germany: Springer. doi:10.1007/3-540-39967-4_13

Bannard, C., Baldwin, T., & Lascarides, A. (2003). A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, (pp. 65–72). Retrieved from http://www.aclweb.org/anthology/W03-1809.

Bonin, F., Dell'Orletta, F., Venturi, G., & Montemagni, S. (2010). Contrastive filtering of domain-specific multi-word terms from different types of corpora. In *Proceedings of the 2010 Workshop on Multiword Expressions: From Theory to Applications*, (pp. 77–80). Retrieved from http://www.aclweb.org/anthology/W10-3711.

Bouma, G. (2010). Collocation extraction beyond the independence assumption. In *Proceedings of the ACL 2010 Conference Short Papers*, (pp. 109–114). Retrieved from http://www.aclweb.org/anthology/P10-2020.

Brunzel, M. (2008). The XTREEM methods for ontology learning from web documents. In Buitelaar, P. (Eds.), *Ontology Learning and Population: Bridging the Gap Between Text and Knowledge* (pp. 3–28). Amsterdam, Netherlands: IOS Press.

Cabré, M. T. (1999). *Terminology: Theory, methods and applications*. Amsterdam, Netherlands: John Benjamins Publishing Company.

Cimiano, P. (2006). *Ontology learning and population from text: Algorithms, evaluation and application*. New York, NY: Springer.

Cucerzan, S. (2007). Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, (pp. 708–716). Retrieved from http://www.aclweb.org/anthology/D/D07/D07-1074.

Cui, G., Lu, Q., Li, W., & Chen, Y. (2008). Corpus exploitation from Wikipedia for ontology construction. In N. Calzolari, et al. (Eds.), *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008),* (pp. 2125–2132). Paris, France: ELRA. Retrieved from http://www.lrec-conf.org/proceedings/lrec2008/pdf/541_paper.pdf.

Da Silva, J. F., & Lopes, G. P. (1999). *A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora.* Paper presented at the 6th Meeting on Mathematics of Language. Orlando, FL.

Erdmann, M., Nakayama, K., Hara, T., & Nishio, S. (2008). An approach for extracting bilingual terminology from Wikipedia. In Haritsa, J. R., Kotagiri, R., & Pudi, V. (Eds.), *DASFAA 2008* (*Vol. 4947*, pp. 380–392). Lecture Notes in Computer Science Berlin, Germany: Springer. doi:10.1007/978-3-540-78568-2_28

Erdmann, M., Nakayama, K., Hara, T., & Nishio, S. (2009). Improving the extraction of bilingual terminology from Wikipedia. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *5*(4), 1–17. doi:10.1145/1596990.1596995

Evert, S., & Krenn, B. (2001). Methods for the qualitative evaluation of lexical association measures. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, (pp. 188–195). Retrieved from http://www.aclweb.org/anthology/P01-1025.

Fazly, A., & Stevenson, S. (2007). Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In *Proceedings of the ACL 2007 Workshop on a Broader Perspective on Multiword Expressions*, (pp. 9–16). Retrieved from http://www.aclweb.org/anthology/W/W07/W07-1102.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*, (pp. 363–370). Retrieved from http://www.aclweb.org/anthology/P/P05/P05-1045.pdf.

Frantzi, K., & Ananiadou, S. (1999). The C-value/NC-value domain independent method for multi-word term extraction. *Journal of Natural Language Processing*, *6*(3), 145–179.

Frantzi, K. T., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms: The C-value/NC-value method. *International Journal on Digital Libraries*, *3*(2), 115–130. doi:10.1007/s007999900023

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, *138*(15), 900–901. doi:10.1038/438900a

Gómez-Pérez, A., Fernández-López, M., & Corcho, O. (2004). *Ontological engineering*. London, UK: Springer.

Hoang, H. H., Kim, S. N., & Kan, M.-Y. (2009). A re-examination of lexical association measures. In *Proceedings of the ACL 2009 Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, (pp. 31–39). Retrieved from http://www.aclweb.org/anthology/W/W09/W09-290.

Kageura, K., & Umino, B. (1996). Methods of automatic term recognition: A review. *Terminology*, *3*(2), 259–289. doi:10.1075/term.3.2.03kag

Katz, G., & Giesbrecht, E. (2006). Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the ACL 2006 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, (pp. 12–19). Retrieved from http://www.aclweb.org/anthology/W/W06/W06-1203.

Kim, J.-D., Ohta, T., Tateisi, Y., & Tsujii, J. (2003). Genia corpus– Semantically annotated corpus for bio-textmining. *Bioinformatics (Oxford, England)*, *19*(1), 1180–1182. doi:10.1093/bioinformatics/btg1023

Korkontzelos, I., Klapaftis, I., & Manandhar, S. (2008). Reviewing and evaluating automatic term recognition techniques. In Ranta, A., & Nordström, B. (Eds.), *Lecture Notes in Artificial Intelligence: GoTAL 2008* (*Vol. 5221*, pp. 248–259). Berlin, Germany: Springer. doi:10.1007/978-3-540-85287-2_24

Korkontzelos, I., & Manandhar, S. (2009). Detecting compositionality in multi-word expressions. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, (pp. 65–68). Retrieved from http://www.aclweb.org/anthology/P/P09/P09-2017.

Kozakov, L., Park, Y., Fin, T., Drissi, Y., Doganata, Y., & Cofino, T. (2004). Glossary extraction and utilization in the information search and delivery system for IBM technical support. *IBM Systems Journal*, *43*(3), 546–563. doi:10.1147/sj.433.0546

Krenn, B., & Evert, S. (2001). Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL Workshop on Collocations*, (pp. 39–46).

Krippendorf, K. (1980). *Content analysis*. London, UK: SAGE Publications.

Kulick, S., Bies, A., Liberman, M., Mandel, M., Mcdonald, R., & Palmer, M. … White, P. (2004). Integrated annotation for biomedical information extraction. In *Proceedings of the HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases,* (pp. 61–68). Retrieved from http://www.aclweb.org/anthology/W/W04/W04-3111.pdf.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. doi:10.2307/2529310

Lenat, D. B. (1995). CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, *38*(11), 33–38. doi:10.1145/219717.219745

Medelyan, O., Milne, D., Legg, C., & Witten, I. H. (2009). Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, *67*(9), 716–754. doi:10.1016/j.ijhcs.2009.05.004

Nakagawa, H., & Mori, T. (1998). Nested collocation and compound noun for term extraction. In *Proceedings of the First Workshop on Computational Terminology (Computerm 1998)*, (pp. 64–70).

Pazienza, M. T., Pennacchiotti, M., & Zanzotto, F. (2005). Terminology extraction: An analysis of linguistic and statistical approaches. In Sirmakessis, S. (Ed.), *Knowledge Mining Series: Studies in Fuzziness and Soft Computing* (pp. 255–279). Heidelberg, Germany: Springer. doi:10.1007/3-540-32394-5_20

Pecina, P., & Schlesinger, P. (2006). Combining association measures for collocation extraction. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, (pp. 651–658). Retrieved from http://www.aclweb.org/anthology/P/P06/P06-2084.

Richman, A. E., & Schone, P. (2008). Mining wiki resources for multilingual named entity recognition. In *Proceedings of ACL 2008: HLT*, (pp. 1–9). Retrieved from http://www.aclweb.org/anthology/P/P08/P08-1001.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In Gelbukh, A. (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 189–206). Berlin, Germany: Springer. doi:10.1007/3-540-45715-1_1

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620. doi:10.1145/361219.361220

Sclano, F., & Velardi, P. (2007). *TermExtractor: A web application to learn the common terminology of interest groups and research communities.* Paper presented at TIA 2007. Sofia Antipolis, France.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). Yago: A core of semantic knowledge. In *Proceedings of the16th international World Wide Web conference (WWW 2007)*, (pp. 697–706). ACM. Retrieved from http://doi.acm.org/10.1145/1242572.1242667.

Toral, A., & Muñoz, R. (2006). A proposal to automatically build and maintain gazetteers for named entity recognition by using Wikipedia. In *Proceedings of the ACL 2006 Workshop on New Text: Wikis and Blogs and Other Dynamic Text Sources*, (pp. 56–61). Retrieved from http://www.aclweb.org/anthology/W/W06/W06-2809.pdf.

Tratz, S., & Hovy, E. (2010). A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, (pp. 678–687). Retrieved from http://www.aclweb.org/anthology/P10-1070.

Villada Moirón, B., & Tiedemann, J. (2006). Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the EACL 2006 Workshop on Multiword Expressions in a Multilingual Context*, (pp. 33–40). Retrieved from http://www.aclweb.org/anthology/W/W06/W06-2405.pdf.

Wermter, J., & Hahn, U. (2005). Finding new terminology in very large corpora. In P. Clark & G. Schreiber (Eds.), *Proceedings of the 3rd International Conference on Knowledge Capture (K-CAP 2005),* (pp.137–144). Retrieved from http://doi.acm.org/10.1145/1088622.1088648.

Zesch, T., Gurevych, I., & Mühlhäuser, M. (2007). Analyzing and accessing Wikipedia as a lexical semantic resource. In Rehm, G., Witt, A., & Lemnitzer, L. (Eds.), *Data Structures for Linguistic Resources and Applications* (pp. 197–205). Tübingen, Germany: Gunter Narr.

Zesch, T., Müller, C., & Gurevych, I. (2008). Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In N. Calzolari, et al. (Eds.), *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008),* (pp. 1646–1652). Paris, France: ELRA. Retrieved from http://www.lrec-conf.org/proceedings/lrec2008/pdf/420_paper.pdf.

Zhang, Z., Iria, J., Brewster, C., & Ciravegna, F. (2008). A comparative evaluation of term recognition algorithms. In N. Calzolari, et al. (Eds.), *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008),* (pp. 2208–2013). Paris, France: ELRA. Retrieved from http://www.lrec-conf.org/proceedings/lrec2008/pdf/538_paper.pdf.

Zhang, Z., Xia, L., Greenwood, M. A., & Iria, J. (2009). Too many mammals: Improving the diversity of automatically recognized terms. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing 2009 (RANLP 2009)*, (pp. 490–495). Retrieved from http://www.aclweb.org/anthology/R09-1087.

ADDITIONAL READING

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*, 555–596. doi:10.1162/coli.07-034-R2

Baldwin, T., Bannard, C., Tanaka, T., & Widdows, D. (2003). An empirical model of multiword expression decomposability. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, (pp. 89–96). Retrieved from http://www.aclweb.org/anthology/W03-1812.

Baldwin, T., & Kim, S. N. (2009). Multiword expressions. In Indurkhya, N., & Damerau, F. J. (Eds.), *Handbook of Natural Language Processing* (2nd ed., pp. 267–292). Boca Raton, FL: CRC Press.

Bunescu, R. C., & Pasca, M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics,* (pp. 9–16). Retrieved from http://www.aclweb.org/anthology/E/E06/E06-1002.pdf.

Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., MacLeod, C., & Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. In M. G. Rodríguez & C. P. S. Araujo (Eds.), *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002*), (pp. 1934–1940). Paris, France: ELRA.

Caseli, H., Villavicencio, A., Machado, A., & Finatto, M. J. (2009). Statistically-driven alignment-based multiword expression identification for technical domains. In *Proceedings of the ACL 2009 Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, (pp. 1–8). Retrieved from http://www.aclweb.org/anthology/W/W09/W09-2901.

De Melo, G., & Weikum, G. (2010). Providing multilingual, multimodal answers to lexical database queries. In N. Calzolari, et al. (Eds.), *Proceedings of the 7th International Conference on Language Resources and Evaluation,* (pp. 348–355). Paris, France: ELRA. Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/pdf/312_Paper.pdf.

Guarino, N. (1998). Formal ontology in information systems. In N. Guarino (Ed.), *1st International Conference on Formal Ontology in Information Systems (FOIS 1998),* (pp. 3–15). Amsterdam, Netherlands: IOS Press.

Guarino, N., Oberle, D., & Staab, S. (2009). What is an ontology? In Staab, S., & Studer, R. (Eds.), *Handbook on Ontologies* (pp. 1–7). Berlin, Germany: Springer. doi:10.1007/978-3-540-92673-3_0

Gurevych, I., & Wolf, E. (2010). Expert-built and collaboratively constructed lexical semantic resources. *Language and Linguistics Compass*, *4*(11), 1074–1090. doi:10.1111/j.1749-818X.2010.00251.x

Justeson, J. S., & Katz, S. M. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, *1*(1), 9–27. doi:10.1017/S1351324900000048

Kageura, K., Daille, B., Nakagawa, H., & Chien, L. F. (2004). Recent trends in computational terminology. *Terminology*, *10*(1), 1–21. doi:10.1075/term.10.1.02kag

Kazama, J., & Torisawa, K. (2008). Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In *Proceedings of ACL-08: HLT*, (pp. 407–415). Retrieved from http://www.aclweb.org/anthology/P/P08/P08-1047.

Kim, S. N., & Kan, M.-Y. (2009). Re-examining automatic keyphrase extraction approaches in scientific articles. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, (pp. 9–16). Retrieved from http://www.aclweb.org/anthology/W/W09/W09-2902.

Maedche, A., & Staab, S. (2004). Ontology learning. In Studer, R., & Staab, S. (Eds.), *Handbook of Information Systems* (pp. 173–190). Berlin, Germany: Springer.

Pecina, P. (2008): A machine learning approach to multiword expression extraction. In *Proceedings of Towards a Shared Task for Multiword Expressions (MWE 2008),* (pp. 54–57). Retrieved from http://www.lrec-conf.org/proceedings/lrec2008/workshops/W20_Proceedings.pdf.

Ramisch, C., Schreiner, P., Idiart, M., & Villavicencio, A. (2008). An evaluation of methods for the extraction of multiword expressions. In *Proceedings of Towards a Shared Task for Multiword Expressions (MWE 2008),* (pp. 50–53). Retrieved from http://www.lrec-conf.org/proceedings/lrec2008/workshops/W20_Proceedings.pdf.

Ruiz-Casado, M., Alfonseca, E., Okumura, M., & Castells, P. (2008). Information extraction and semantic annotation of Wikipedia. In Buitelaar, P. (Eds.), *Ontology Learning from Text: Methods, Evaluation and Applications* (pp. 91–106). Amsterdam, Netherlands: IOS Press.

Schone, P., & Jurafsky, D. (2001).Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, (pp. 100–108). Retrieved from http://www.aclweb.org/anthology-new/w/w01/w01-0513.pdf.

Villavicencio, A., Bond, F., Korhonen, A., & McCarthy, D. (2005). Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Computer Speech & Language*, *19*(4), 365–377. doi:10.1016/j.csl.2005.05.001

Zesch, T. (2010). What's the difference? Comparing expert-built and collaboratively-built lexical semantic resources. In N. Calzolari, P. Baroni, M. Monachini, & C. Soria (Eds.), *Proceedings of the 2nd European Language Resources and Technologies Forum Language Resources of the Future / the Future of Language Resources,* (pp. 91–92). Retrieved from http://www.flarenet.eu/sites/default/files/FLaReNet_Forum_2010_Proceedings.pdf.

KEY TERMS AND DEFINITIONS

Collocation: A collocation is a type of multiword expression. Collocations are semantically compositional, but are lexicalized terms due to being typically used as a unit to refer to a particular concept.

Multiword Expression: A lexical unit in general language consisting of more than one word. Multiword expressions receive their own entry in a natural language lexicon because of their irregular semantic, syntactic, pragmatic or statistical properties.

Multiword Term: A term consisting of more than one word.

Semantic Compositionality: Semantic compositionality is a property of multiword expressions. If the meaning of a multiword expression can be inferred from the meaning of its constituent words, the multiword expression is semantically compositional. If this is only partly or not at all possible, the multiword expression shows weak or strong non-compositionality.

Term: A lexical unit (a word or a phrase) representing a domain-relevant concept.

Termhood: Termhood refers to the degree of domain relevance of a lexical unit to a particular domain. Domain-specific terms have high termhood.

Unithood: Unithood refers to how strong the words in a phrase are associated with each other to form a lexical unit. Multiword expressions have high unithood.