Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computer Speech and Language 18 (2004) 275–299
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Unsupervised and supervised exploitation of semantic domains in lexical disambiguation ☆

Alfio Gliozzo [a], Carlo Strapparava [a,*], Ido Dagan [b]

[a] *ITC-irst, Istituto per la Ricerca Scientifica e Tecnologica, 1-38050 Trento, Italy*
[b] *Department of Computer Science, Bar Ilan University, Ramat Gan, Israel*

Received 3 October 2003; received in revised form 12 May 2004; accepted 20 May 2004
Available online 9 June 2004

Abstract

Domains are common areas of human discussion, such as economics, politics, law, science, etc., which are at the basis of lexical coherence. This paper explores the dual role of domains in word sense disambiguation (WSD). On one hand, domain information provides generalized features at the paradigmatic level that are useful to discriminate among word senses. On the other hand, domain distinctions constitute a useful level of coarse grained sense distinctions, which lends itself to more accurate disambiguation with lower amounts of knowledge.

In this paper we extend and ground the modeling of domains and the exploitation of WORDNET DOMAINS, an extension of WORDNET in which each synset is labeled with domain information. We propose a novel unsupervised probabilistic method for the critical step of estimating domain relevance for contexts, and suggest utilizing it within unsupervised domain driven disambiguation for word senses, as well as within a traditional supervised approach.

The paper presents empirical assessments of the potential utilization of domains in WSD at a wide range of comparative settings, supervised and unsupervised. Following the dual role of domains we report experiments that evaluate both the extent to which domain information provides effective features for WSD, as well as the accuracy obtained by WSD at domain-level sense granularity. Furthermore, we demonstrate the potential for either avoiding or minimizing manual annotation thanks to the generalized level of information provided by domains.

☆ This work was developed under the collaboration ITC-irst/University of Haifa.

* Corresponding author.
E-mail addresses: gliozzo@itc.it (A. Gliozzo), strappa@itc.it (C. Strapparava), dagan@cs.biu.ac.il (I. Dagan).

0885-2308/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2004.05.006

1. Introduction and motivations

Domains are common areas of human discussion, such as economics, politics, law, science, etc. (see Table 1), which demonstrate lexical coherence. A substantial portion of the language terminology may be characterized as *domain words* whose meaning refers to concepts belonging to specific domains, and which often occur in texts that discuss the corresponding domain.

Domains have been used with a dual role in linguistic description. One role is characterizing word senses, typically as the semantic field of a word sense in a dictionary or lexicon (e.g. *crane* has senses in the domains of ZOOLOGY and CONSTRUCTION). The WORDNET DOMAINS lexical resource is an extension of WORDNET which provides such domain labels for all synsets (Magnini and Cavaglià, 2000). A second role is to characterize texts, typically as a generic level of text categorization (e.g. for classifying news and articles) (Sebastiani, 2002).

From the perspective of word sense disambiguation domains may be considered from two points of view. First, a major portion of the information required for sense disambiguation corresponds to paradigmatic domain information. Many of the features that contribute to disambiguation identify the domains that characterize a particular sense or subset of senses. For example, economics terms provide characteristic features for the financial senses of words like *bank* and *interest*, while legal terms characterize the judicial sense of *sentence* and *court*. Common supervised WSD methods capture such domain-related distinctions separately for each sense of each word, and may require relatively many training examples in order to obtain sufficiently many features of this kind for each sense (Yarowsky and Florian, 2002). However, domains represent an independent linguistic notion of discourse, which does not depend on a specific word sense. Therefore, it is beneficial to model a relatively small number of domains directly, as a generalized notion, and then use the same generalized information for many instances of the WSD task. A major goal of this paper is to study the extent to which domain information can contribute along this vein to WSD.

Table 1
Domain distribution over WORDNET synsets

Domain	#Syn	Domain	#Syn	Domain	#Syn
Factotum	36820	Biology	21281	Earth	4637
Psychology	3405	Architecture	3394	Medicine	3271
Economy	3039	Alimentation	2998	Administration	2975
Chemistry	2472	Transport	2443	Art	2365
Physics	2225	Sport	2105	Religion	2055
Linguistics	1771	Military	1491	Law	1340
History	1264	Industry	1103	Politics	1033
Play	1009	Anthropology	963	Fashion	937
Mathematics	861	Literature	822	Engineering	746
Sociology	679	Commerce	637	Pedagogy	612
Publishing	532	Tourism	511	Computer_Science	509
Telecommunication	493	Astronomy	477	Philosophy	381
Agriculture	334	Sexuality	272	Body_Care	185
Artisanship	149	Archaeology	141	Veterinary	92
Astrology	90				