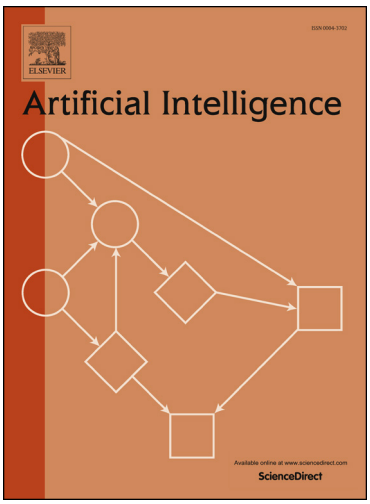
Accepted Manuscript



NASARI: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities

José Camacho-Collados, Mohammad Taher Pilehvar, Roberto Navigli

PII: S0004-3702(16)30082-0
DOI: http://dx.doi.org/10.1016/j.artint.2016.07.005
Reference: ARTINT 2964

To appear in: *Artificial Intelligence*

Received date: 23 December 2015
Revised date: 14 July 2016
Accepted date: 25 July 2016

Please cite this article in press as: J. Camacho-Collados et al., NASARI: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities, *Artif. Intell.* (2016), http://dx.doi.org/10.1016/j.artint.2016.07.005

NASARI: Integrating Explicit Knowledge and Corpus Statistics for a Multilingual Representation of Concepts and Entities

José Camacho-Collados, Mohammad Taher Pilehvar[a,1], Roberto Navigli

Department of Computer Science
Sapienza University of Rome

[a] *Department of Theoretical and Applied Linguistics*
University of Cambridge

Abstract

Owing to the need for a deep understanding of linguistic items, semantic representation is considered to be one of the fundamental components of several applications in Natural Language Processing and Artificial Intelligence. As a result, semantic representation has been one of the prominent research areas in lexical semantics over the past decades. However, due mainly to the lack of large sense-annotated corpora, most existing representation techniques are limited to the lexical level and thus cannot be effectively applied to individual word senses. In this paper we put forward a novel multilingual vector representation, called NASARI, which not only enables accurate representation of word senses in different languages, but it also provides two main advantages over existing approaches: (1) high coverage, including both concepts and named entities, (2) comparability across languages and linguistic levels (i.e., words, senses and concepts), thanks to the representation of linguistic items in a single unified semantic space and in a joint embedded space, respectively. Moreover, our representations are flexible, can be applied to multiple applications and are freely available at `http://lcl.uniroma1.it/nasari/`. As evaluation benchmark, we opted for four different tasks, namely, word similarity, sense clustering, domain labeling, and Word Sense Disambiguation, for each of which we report state-of-the-art performance on several standard datasets across different languages.

Keywords:
semantic representation, lexical semantics, Word Sense Disambiguation, semantic similarity, sense clustering, domain labeling

1. Introduction

Semantic representation, i.e., modeling the semantics of a linguistic item[2] in a mathematical or machine interpretable form, is a fundamental problem in Natural Language Processing (NLP) and Artificial Intelligence (AI). Because they represent the lowest linguistic level, word senses play a vital role in natural language understanding. Effective representations of word senses can be directly useful to Word Sense Disambiguation [94], semantic similarity [13, 130, 107], coarsening sense inventories [93, 125], alignment of lexical resources [102, 99, 109], lexical substitution [75], and semantic priming [101]. Moreover, sense-level representation can be directly extended to applications requiring word representations, with the added benefit that it provides extra semantic information. Turney and Pantel [130] provide a review of some of the applications of word representation, including: automatic thesaurus generation [21, 22], word similarity [25, 129, 114] and clustering [104], query expansion [141], information extraction [61], semantic role labeling [29, 105], spelling correction [53], and Word Sense Disambiguation [94].

[1] Work mainly done at the Sapienza University of Rome.

[2] Throughout this article by a linguistic item we mean any kind of linguistic unit that can bear a meaning, i.e., a word sense, a word, a phrase, a sentence or a larger piece of text.