Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language xxx (2016) xxx–xxx
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

# Integrated concept blending with vector space models[☆]

Q1 Hiram Calvo[*], Oscar Méndez, Marco A. Moreno-Armendáriz

*Centro de Investigación en Computación, Instituto Politécnico Nacional, Av. Juan de Dios Bátiz s/n, México D.F. 07738, Mexico*

Received 19 August 2014; received in revised form 23 December 2015; accepted 29 January 2016

## Abstract

Traditional concept retrieval is based on usual word definition dictionaries with simple performance: they just map words to their definitions. This approach is mostly helpful for readers and language students, but writers sometimes need to find a word that encompasses a set of ideas that they have in mind. For this task, inverse dictionaries are ready to help; however, in some cases a sought word does not correspond to a single definition but to a composite meaning of several concepts. A language producer then tends to require a concept search that starts with a group of words or a series of related terms, looking for a target word. This paper aims to assist on this task by presenting a new approach for concept blending through the development of a search-by-concept method based on vector space representation using semantic analysis and statistical natural language processing techniques. Words are represented as numeric vectors based on different semantic similarity measures and probabilistic measures; the semantic properties of a word are captured in the vector elements determined by a given linguistic context. Three different sources are used as context for word vector construction: WordNet, a distributional thesaurus, and the Latent Dirichlet Allocation algorithm; each source is used for building a different semantic vector space.

The concept-blender input is then conformed by a set of n-nouns. All input members are read and substituted by their corresponding vectors. Then, a semantic space analysis including a filtering and ranking process is carried out to deploy a list of target words. A test set of 50 concepts was created in order to evaluate the system's performance. A group of 30 evaluators found our integrated concept blending model to provide better results for finding an adequate word for the provided set of concepts.

*Keywords:* Computational linguistics; Natural language processing; Lexicography; Vector space models; Reverse lookup dictionaries; Concept-blending

## 1. Introduction

Q3 Generally, traditional dictionaries are designed with readers in mind. In this scenario, a query is based on looking up single words in order to get their corresponding meanings; this is not always helpful when viewed from language producer's perspective. For them –speakers, writers, etc.– necessities are different: they may have an idea composed by several meanings or concepts, and their goal is to find the best word that is able to represent their thoughts. As

[☆] This paper has been recommended for acceptance by Roger K. Moore.

Q2 [*] Corresponding author. Tel.: +52 15557296000.
*E-mail addresses:* hcalvo@cic.ipn.mx (H. Calvo), omendez_a12@cic.ipn.mx (O. Méndez), mam_armendariz@cic.ipn.mx (M.A. Moreno-Armendáriz).

http://dx.doi.org/10.1016/j.csl.2016.01.004
0885-2308/© 2016 Published by Elsevier Ltd.

more dictionaries in electronic format are available, searching by concept is readily at hand; but, as we will detail in Section 2, current implementations have limitations due to the presence of syntactic problems, query expansion issues, and gloss reliance. The implementation of vector-space word representations enables circumventing most of these well-known problems through their capacity to capture syntactic and semantic regularities in language (Mikolov et al., 2013). Also, working with continuous space models permits a distributed representation with good levels of generalization; moreover, semantic vector spaces have the characteristic that similar words tend to have similar vectors.

This paper presents a new approach for concept retrieval, based on vector space representation constructed with semantic similarity measures and statistical NLP techniques. A sought concept is expressed as an input of $n$ nouns. We propose representing each noun as a numeric vector in order to allocate them in a so-called semantic space of $m$ dimensions; these dimensions correspond to a predefined set of concepts, words or topics. Then, the value of each element in the noun's vector is given by its relation (similarity measure, or probabilistic distribution with regard to a topic) with those reference concepts, words or topics. Given the input set of points represented in the $m$-dimensional vector space, we find an equidistant new point, from which a sample of the nearest neighbors is taken. The words included in this sample should semantically mix the characteristics described by the original entries, representing the target words from the reverse lookup process of our concept blending method.

The semantic space is created from three different sources in order to evaluate different approaches: a supervised approach assisted by WordNet, and two unsupervised approaches using (1) a distributional thesaurus, and (2) the Latent Dirichlet Allocation (LDA) algorithm.

Once the concept blending method was ready, a test set was created to attest the performance of our proposal, and an evaluation procedure determined which one of the proposed models used for the semantic space creation was closer to human associative reasoning. Results were compared with an existing search-by-concept dictionary (OneLook), showing that our results were preferred, in the majority of cases, by master class evaluators. Additionally, we found a greater semantic association for the words provided by our method.

In Section 2 we present similar works in the state of the art; in Section 3 we present our method. In Section 4 our experiments and results can be found, and finally, our conclusions are drawn.

## 2. State of the art

This work uses the representation of words as a vector. This is not a new idea, perhaps the earliest works on this were those of Hinton et al. (1986) and Hodgson (1991). In Section 2.1 we give a brief survey on this subject, while in Section 2.2 we present works related to concept retrieval.

### *2.1. Distributional semantics models*

Distributional analysis is based in structuralist linguistics (Harris, 1951), corpus linguistics (Firth, 1968), psychology (Miller and Charles, 1991), and it is based on the idea that not only the semantic properties of a lexical item are fully reflected in appropriate aspects of the relations it contracts with actual and potential contexts, but “there are good reasons for a principled limitation to linguistic contexts” Cruse (1986). Distributional hypothesis suggests that we can induce aspects of the meaning of words from their contexts; it is a “theory of meaning” that can be easily operationalized into a procedure to extract “meaning” from text corpora on a large scale.

Models that represent the meaning of words as vectors keeping track of the words' distributional history focus on the notion of semantic similarity, measured with geometrical methods in the space inhabited by the distributional vectors. Examples of these models are LSA (Landauer and Dumais, 1997), HAL (Lund and Burgess, 1996), the work of Sahlgren (2006), Padó and Lapata (2007), and Baroni and Lenci (2010). We follow the principle of distributional semantics as models of word meaning following Landauer and Dumais (1997), and Turney et al. (2010).

Distributional semantics can model human similarity judgments, lexical priming (*hospital* primes *doctor*) synonymy (*zenith-pinnacle*), analogy (*mason* is to *stone* like *carpenter* is to *wood*), relation classification (*exam-anxiety*:CAUSE-EFFECT), etc. One of the earliest works is Hodgson's (1991). He found similar amounts of priming for different semantic relations between primes and targets (approx. 23 pairs per relation), for example: synonyms: to dread/to fear, antonyms: short/tall, coordinates: train/truck, super- and subordinate pairs: container/bottle, free association pairs: dove/peace, phrasal associates: vacant/building.