A Novel Distributional Approach to Multilingual Conceptual Metaphor Recognition

Michael Mohler and **Bryan Rink** and **David Bracewell** and **Marc Tomlinson**
Language Computer Corp.
Richardson, Texas, USA
`{michael,bryan,david,marc}@languagecomputer.com`

Abstract

We present a novel approach to the problem of multilingual conceptual metaphor recognition. Our approach extends recent work in conceptual metaphor discovery by combining a complex methodology for facet-based concept induction with a distributional vector space model of linguistic and conceptual metaphor. In the evaluation of our system in English, Spanish, Russian, and Farsi, we experiment with several state-of-the-art vector space models and demonstrate a clear benefit to the fine-grained concept representation that forms the basis of our methodology for conceptual metaphor recognition.

1 Introduction

The role of metaphor in language has been defined by Lakoff et al. (1980; 1993) as a cognitive phenomenon which operates at the level of mental processes, whereby one concept or domain is viewed systematically in terms of another. For example, the phrase "to cure poverty" is a metaphor which subtly conveys a wide variety of information to the listener. In order to mentally process this phrase, we must first recognize that a metaphor is being used and that "cure" (as a medical term) is being used figuratively. Then, we assume some relationship between "poverty" and "things that can be medically cured" which leads to the conceptual mapping "POVERTY as DISEASE." This conceptual mapping enables the listener to transfer a variety of properties and associations between the two concepts, such as their association with a feeling of helplessness, the existence of sustained efforts to end them, the potential for them to spread, and their mutual relationship with ill-health and death. Therefore, by identifying the conceptual domains associated with this linguistic metaphor, we are able to reason about the target domain (POVERTY) using concepts and terms associated with the source domain (DISEASE).

Any natural language processing system capable of processing metaphor in text with human-level competence must, therefore, overcome three problems in sequence:

1. the identification of metaphorical expressions (also known as linguistic metaphors (LMs))

2. the discovery of a conceptual domain mapping or conceptual metaphor (CM) which consists of

    (a) the conceptual domain of the metaphor target (e.g., POVERTY); and
    (b) the conceptual domain of the metaphor source (e.g., DISEASE)

3. the real-world interpretation of the metaphorical text which uses the conceptual metaphor framework to transfer knowledge between the source and target domains.

While a significant amount of recent work has presented interesting and promising methodologies for multilingual LM identification (Shutova and Sun, 2013; Wilks et al., 2013; Strzalkowski et al., 2013), the work presented in this paper is focused on (2), the problem of multilingual CM recognition, which will be made to serve as the foundation for a more fine-grained interpretation of metaphor.

We cast the CM recognition process as a two-part methodology which (a) selects the target domain associated with a particular LM that has been detected; and (b) determines the source domain to which it should be mapped in order to produce a satisfactory interpretation. In this work, we assume that the target domains are known and belong to one of the following conceptual spaces: POVERTY, WEALTH, or TAXATION. Pragmatically speaking, research in CM recognition presupposes some methodology for LM identification, and to this end, we have employed an existing state-of-the-art LM identification system which has been developed to detect linguistic metaphors in four languages: English, Spanish, Russian, and Farsi (Bracewell et al., 2014).

In order to generate a CM which can serve as the basis for an interpretation of an LM, we have developed an approach that is based on the following hypotheses:

CONCEPTUAL HYPOTHESIS: When an LM has been identified as a pair of lexical items that represent the source (e.g., "cure") and the target (e.g., "poverty"), we can generate a conceptual mapping by selecting the conceptual domains that are, *a priori*, the most likely for the source and target lexemes.[1]

DISTRIBUTIONAL HYPOTHESIS: It is possible to decide which conceptual space better represents a given lexeme by

1. expanding the lexical space with additional terms (which we call "grammatical co-occurrents") that are strongly associated with the lexeme through grammatical relations such as AGENT, PATIENT, INSTRUMENT, and ATTRIBUTE;
2. using these lexical expansions to produce distributional vectors; and
3. uncovering the *selectional constraints* of particular domain facets by clustering the distributional vectors within a semantic space.

DOMAIN HYPOTHESIS: The grammatical co-occurrents of the LM are themselves very likely to belong to the same conceptual domain as the lexeme (e.g., "cure patient", "cured of AIDS", and "doctor cured").

MAPPING HYPOTHESIS: The semantic space representations of both the LM source and its grammatically associated terms can be used to produce mappings into a high dimensional space in which source domains are known to exist.

While other computational linguistics research in metaphor has made use of the CONCEPTUAL and DISTRIBUTIONAL hypotheses, to our knowledge the DOMAIN and MAPPING hypotheses have not yet been explored in combination with a distributional approach.

The remainder of this work is organized as follows. In Section 2, we discuss related work in the field of metaphor interpretation and unsupervised concept induction. In Section 3, we introduce the overall architecture of our CM recognition system. In Section 4, we describe our method for representing lexical items and conceptual metaphors in a distributional vector space. Then, in Section 5, we explain our methodology for creating and ranking clusters of LM co-occurrents which are then mapped to conceptual metaphors within our vector space. In Section 6, we describe our experimental setup and provide the results of our experiments. Finally, in Section 7 we present our conclusions.

2   Related Work

Research in metaphor processing can broadly be divided into two categories: metaphor identification and metaphor interpretation. Although some recent work on metaphor interpretation has skirted the issue of conceptual metaphor entirely by casting the problem of metaphor interpretation as an instance of lexical paraphrase (Shutova, 2010; Bollegala and Shutova, 2013) or textual entailment (Mohler et al., 2013), the mapping and modeling of conceptual metaphors has historically served as an important foundation for

---

[1]If the target domains are pre-selected, this hypotheses is reduced to selecting only the most likely source domain.

more robust interpretation of metaphor. Indeed, a significant amount of research in metaphor interpretation has been concentrated on the development of highly-structured, manually curated representations of both the CM source and CM target domains. Notable in this regard are the KARMA system (Feldman and Narayanan, 2004) which was designed to simulate neurological modeling of verbs – both abstract and metaphorical – and the ISOMETA system (Beust et al., 2003) which made use of differential tables of CM domain lexical items to drive their metaphor interpretation process. The CorMet system (Mason, 2004) sought to model conceptual metaphors by detecting individual source-target mappings that provide evidence for a known CM by quantifying the overlap between clusters of terms with a strong selectional preference to the most representative verbs within the source and target domains. After a manual inspection of the source/target cluster pairs across domains, the directionality and the systematicity of these underlying conceptual mappings were quantified in order to produce an overall confidence in the mapping. As part of their development of the Hamburg Metaphor Database (HMD), Reining and Lönneker-Rodman (2007) performed a a manual categorization of lexical items into conceptual source domains with a facet-level granularity and enriched their domains using a WordNet-based lexical expansion. In the same vein, Chung et al. (2005) chose to model source domains by expanding their lexical items by exploiting the links between WordNet glosses and the SUMO ontology.

In recent years, however, research has focused on automating the modeling and classification of conceptual metaphors as much as possible in order to encourage the scaling up of metaphor research in general. Veale and Hao (2008), as part of the Talking Points system, developed what they refer to as a Slipnet which defines linked chains of meaning that connect a source to a target through shared (or related) attributes and actions. As a step in this process, they combined WordNet relations with pragmatic relations extracted from text and clustered nouns according to their relation (and attribute) similarity in order to define a weak conceptual mapping within the clusters. In a similar way, Shutova et al. (2010), beginning with a seed set of noun/verb linguistic metaphor pairs, performed spectral clustering on a large set of nouns and verbs in order to predict metaphors which participate in the same conceptual metaphor mapping. In particular, she modeled verbs according to their subcategorization frames parameterized by a model of their selectional preferences, while nouns were modeled according to the verbs with which they frequently co-occurred in a dependency relation.

More recently, Gandy et al. (2013) approached the CM discovery problem as a set covering problem. For a given nominal target lexeme, they began by finding all facets (i.e., verbs/adjectives) that share a positive PMI with the target. Then, they would find the set of nouns that also have a positive PMI with those facets, compute their confidence in each association, and heuristically select pairs of concepts (defined as rooted WordNet synset trees) which subsume a large percentage of those nouns and cover a large portion of the overlapping facets. Similarly, Shutova and Sun (2013) detect conceptual mappings by performing hierarchical graph factorization clustering on a graph in which the vertices are defined to be nouns (i.e., concepts) and the edges are weighted using Jensen-Shannon Divergence. For a given input LM source, its likely conceptual metaphors are then discovered by determining its non-literal cluster membership. Finally, Strzalkowski et al. (2013) discovered terms (literal and metaphoric) which often co-occur with an LM source in a corpus and clustered those terms using WordNet and corpus statistics to form "ProtoSources" which could be further inspected to define CM source concepts.

Two vector-based approaches to concept representation are of particular interest in understanding the present work. In the first of these, Schütze (1998) described an approach to word sense identification using second-order co-occurrence vectors which were used to cluster first-order vectors of the in-context terms into senses.[2] Lin (1998), in developing a methodology for evaluating the quality of thesauri, defined a word vector space that moved beyond simple co-occurrence by integrating information about the relations between the word and its co-occurrents. In particular, a word's vector was defined by the number of times that word occurred within a set of (*word, relation, word*) tuples. Our DepVec space represents an extension to Lin's space insofar as we incorporate additional information about relational (i.e., selectional) preference.

---

[2]While context is critical in word sense disambiguation, we hasten to point out that one mark of metaphoricity is its disconnect from the surrounding literal context.
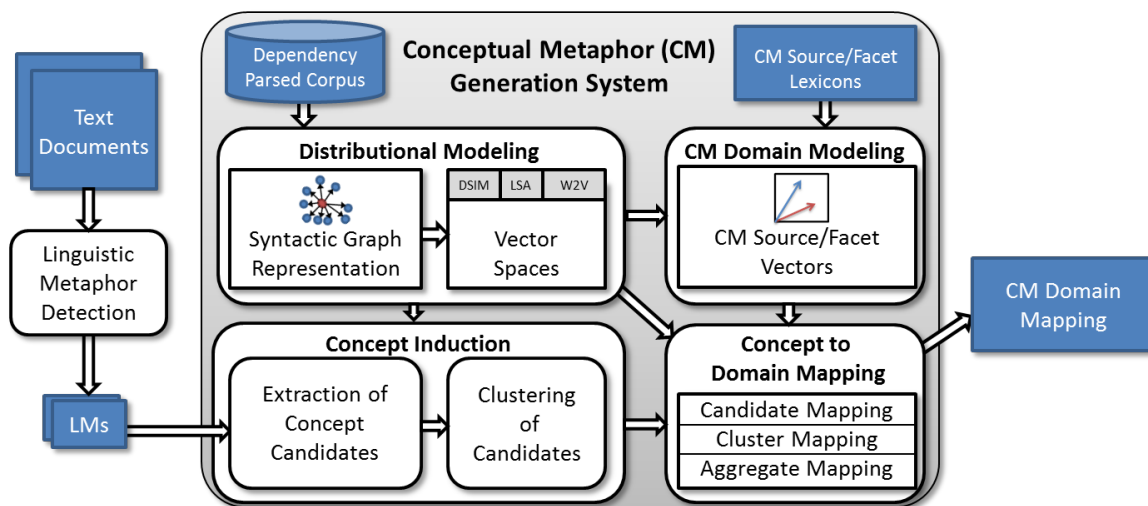
Figure 1: The architecture of our conceptual metaphor recognition system. This system takes a linguistic metaphor as input, induces potential concepts using vector-space clustering, and maps these clusters onto a conceptual metaphor domain.

3 A New Methodology for Conceptual Metaphor Recognition

Figure 1 shows the overall flow of our metaphor processing architecture. We begin with a set of documents gathered from a variety of online news-wire sources. These documents are fed to our state-of-the-art LM detection system which employs a binary logistic regression classifier using a variety of feature modules including imageability and concreteness estimation, topicality modeling, pattern matching, semantic categorization, selectional preference violation, and source/target vector space similarity. The methodology used in this system is beyond the scope of this work, but it is described in detail by Bracewell et al. (2014). The LMs provided by the detection system are validated by a group of native-language experts before being sent for CM recognition system for concept-level interpretation.

Once the LMs have been collected and validated, the CM recognition system begins by extracting, weighting, and clustering the common grammatical contexts of the LM source term. By grammatical context, we refer to the syntactic relations (along with their arguments) which have been found to frequently co-occur with the LM source term in open text. In order to model this grammatical context, we have syntactically parsed a wide collection of documents in each of our focus languages: English, Spanish, Russian, and Farsi. From these parsed documents, we have extracted the most common grammatical co-occurrents of each word in the corpus along with the relation that connects them and the number of times they are connected by that relation. For a given word, we refer to the set of its grammatical co-occurrents as the "concept candidates" associated with that word, as they represent potential concepts within the same conceptual domain as the given word (the DOMAIN HYPOTHESIS). For example, grammatical co-occurrents of the noun "battle" would include many WAR concepts such as "fought", "died in", "waged", "naval", and "losing".

Since a conceptual domain is made up of several interacting concepts, we perform a clustering over the grammatical co-occurrents to produce groups of terms which are likely to represent individual concepts within a domain. The clustering is performed within a high-dimensional, distributional vector space which we describe in Section 4. The clusters are then merged and aligned with a set of 51 pre-defined source concept domains (see Table 1) that have been found to occur frequently in conceptual metaphors about POVERTY, WEALTH, or TAXATION. For each of these known conceptual domains, we have amassed a collection of lexical items for the purpose of modeling the domains and aligning them to our automatically discovered domains. The collection of lexical items associated with each domain have been further partitioned into three to five *facets* which provide a more fine-grained representation of the domain. For instance, the conceptual domain of ABYSS as been subdivided into facets representing

1755

Full Source Concept List					
A GOD	COMPETITION	ENSLAVEMENT	LIGHT	NATURAL PHYSICAL FORCE	PORTAL
A RIGHT	CONFINEMENT	FOOD	LOW POINT	OBESITY	RESOURCE
ABYSS	CRIME	FORCEFUL EXTRACTION	MACHINE	PARASITE	SCHISM
ACCIDENT	CROP	GAME	MAZE	PATHWAY	STRUGGLE
ADDICTION	DARKNESS	GEOGRAPHIC FEATURE	MEDICINE	PHYSICAL BURDEN	VERTICAL SCALE
ANIMAL	DESTROYER	GOAL DIRECTED	MONSTER	PHYSICAL HARM	VISION
BLOOD SYSTEM	DISEASE	HIGH POINT	MORAL DUTY	PHYSICAL LOCATION	
BODY OF WATER	ENABLER	HUMAN BODY	MOVEMENT	PHYSICAL OBJECT	
BUILDING	ENERGY	IMPURITY	MOVEMENT ON A VERTICAL SCALE	PLANT	
Sample Lexical Items					
ANIMAL	bite, bark, claw, bird, beaver		MEDICINE	dosage, prescription, heal	
ENSLAVEMENT	servant, oppression, ruler		STRUGGLE	enemy, fight, combat, attack	

Table 1: The 51 source conceptual domains along with some sample English lexical items for a subset of them.

DEPTH (e.g., "deep", "bottomless"), ENTRANCE (e.g., "plunged into", "falling into"), and EXIT (e.g., "climb out of").

3.1 Motivating Example

Table 2 shows a sample of the concept candidates associated with the word "cure" along with the relation that connects them. Our methodology for extracting these terms is discussed in Section 5.1.

nsubj	NIH, WHO, therapist, doctor, vaccine, drug, medicine, *chef, butcher*	prep_of	cancer, AIDS, HIV, malaria, influenza, seizures, allergies
dobj	cancer, polio, Goji Berries, man, genetic defects, aging, infant, woman, depression, *meat, fish, garlic*	prep_by	bone marrow transplant, spleen cells, acupuncture, *smoking, salting,* doxycycline, *drying, burying, dipping*
prep_without	surgery, operation, suppuration, *salt* chemotherapy, injections	prep_to^{-1}	need, project, *brine, mineral,* coalition, run, walk, *salt, nitrite*
prep_in	mice, children, baby, *spices, salt,* monkeys, drug trial, breakthrough, *brine, smokehouse, basement, fridge*	prep_for	*grinding, smoking, voyages, lox, transportation, preservation, jerky, sausages, bacon, sale*

Table 2: Terms that are frequently a part of the grammatical context of "cure" along with their associated relations

It is clear from the concept candidates shown that there are at least two coarse-grained senses of "cure" present – corresponding to the domains of MEDICINE and FOOD. Table 3 shows a sample result of clustering these concept candidates. These clusters are organized according to their domain with MEDICINE-related clusters in the left grouping, FOOD-related clusters in the top-right grouping, and clusters not strongly related to either domain in the bottom-right grouping. Each row of the table represents a single cluster. In addition, it can be observed that these clusters correspond to particular semantic facets of the conceptual domain. For instance, there is a cluster that defines "procedures which result in medical cures" ("acupuncture", "surgery", "operation", etc.), one that defines "individuals who cure food products" ("chef", "butcher"), and one that defines "diseases that can (potentially) be cured" ("cancer", "polio", "AIDS", etc.). Our methodology for automatically inducing such clusters is described in Section 5.2.

Once the clusters have been identified, they can be used to define a mapping from the original LM ("cure") onto a pre-defined set of CM source domains (the MAPPING HYPOTHESIS). In particular, individual concept candidates are mapped to CM domains by calculating the distance between the candidate and one or more vectors representing each domain in a high-dimensional distributional vector space.

4 Distributional Representations

Our method for identifying conceptual metaphor domains relies on determining when multiple words should be grouped as belonging to the same conceptual class (the DISTRIBUTIONAL HYPOTHESIS). Previous work in semantic similarity has shown two types of approaches to work well: (a) hand-coded knowledge such as WordNet or SUMO, and (b) distributional approaches which rely on statistics of

NIH, WHO, therapist, doctor	chef, butcher
vaccine, drug, medicine, doxycycline	project, coalition
spleen cells, bone marrow transplant	meat, fish, sausages, jerky, bacon, lox
acupuncture, surgery, operation	garlic, Goji Berries
chemotherapy, injections, suppuration	smoking, salting, drying, dipping
HIV, malaria, influenza	burying
cancer, polio, AIDS	salt, brine, spices, nitrite, mineral
genetic defects, aging, depression	smokehouse, basement, fridge
seizures, allergies	run, walk
drug trial, breakthrough	voyages, transportation
infant, man, woman, children, baby	mice, monkeys

Table 3: Terms from Table 2 grouped into conceptual clusters – one per line. These clusters are organized according to their domain association: MEDICINE (left), FOOD (top-right), unclear (bottom-right).

word usage in corpora. We adopt the distributional approach in order to facilitate research in languages (such as Farsi) for which coverage of existing knowledge bases is limited. The only requirements for our approach are a corpus with documents written in that language and a syntactic parser for the language. We use the Malt dependency parser to obtain syntactic parses for web documents in each language.

Table 2 of Section 3.1 shows some of the words which participate regularly with the word "cure" in a dependency relation. These syntactic contexts of the word "cure" form the basis for one semantic representation we use to find other similar words, which we will call *DepVec*. All of the dependency relations for a word are used to form a vector-based distributional representation for that word. This representation projects words which are semantically similar to one another onto vectors which are near to each other in the vector space. In the following subsection, we describe DepVec along with LSA and word2vec which are alternative vector space models of word meaning. These vector spaces are then used to calculate similarities between words in order to cluster them and to align them with lexicons which model our existing conceptual spaces.

4.1 Dependency Vectors (DepVec) space

In our DepVec vector space model, each word is represented by a vector whose elements correspond to syntactic contexts of the word. Each element of the vector for word w corresponds to the frequency of a unique dependency relation (w, r, w') seen in the corpus. For example, if the relation $(whale, nsubj^{-1}, swim)$ is extracted once, then the vector for "whale" contains a 1 in the element for $(nsubj^{-1}, swim)$, and the vector for "swim" contains a 1 for the element $(nsubj, whale)$. This representation corresponds that proposed by Lin (1998).

However, the use of raw frequency counts in these vectors leads to a situation in which words that are more frequent in the corpus (e.g., "of", "the", "one") will have higher frequencies in the vectors by chance alone, and so a high co-occurrence count for those words is not indicative of a significant relation to the word. We overcome this limitation by replacing the raw frequency counts in each vector with their corresponding G-test scores. The G-test is a measure of statistical significance for proportions, similar to the Chi-square test, which measures the degree to which a particular triple (w, r, w') was found to occur more frequently than expected given all relations (w'', r, w'). If w' occurs far more often with w than it does with other words, then it will receive a high G-test score for w. In particular, the G-test score is computed according to the following equation:

$$G = 2 \sum_i O_i \cdot ln(O_i/E_i)$$

where the index i ranges over the four cells of a 2x2 contingency table, O_i is the observed count in cell i, and E_i is the expected count in the same cell.

Language	Source	# Documents	Language	Source	# Documents
English	ClueWeb	13,361,743	Spanish	ClueWeb	3,682,478
Russian	ruWac	1,173,590	Farsi	Online news sites	835,588

Table 4: Statistics of the corpora used to construct the vector space models

4.2 Latent Semantic Analysis (LSA)

While the DepVec model provides information about the immediate contexts a word can be expected to occur in, it does not directly capture information about the broader contexts typical of that word, such as topical information. Latent Semantic Analysis (LSA) is a well-studied model (Landauer and Dumais, 1997) which does capture such topical information. The LSA model utilizes a singular value decomposition of a TF-IDF weighted matrix representation of the term-document co-occurrences. Terms and documents are then represented in a reduced dimensionality space using only the information from the eigenvectors with the k largest eigenvalues.

4.3 Continuous skip-gram model (W2V)

Mikolov et al. (2013) recently presented a new method for determining distributional word representations based on a shallow neural network model. The values of the latent vector for each word are trained to optimize prediction of the words within a 10 token window. This prediction is performed using the term's latent vector as the input to a series of log-linear classifiers with outputs which correspond to probability distributions over the tokens within the context window. Each position in the context window is assigned its own classifier weights, so that the model used for making predictions about words immediately following the input term is different than the model which makes predictions about the words two tokens after the term, and so on. Because these latent vector representations are in a low dimensionality space (300 dimensions in our case), the training process will tend to move the representations for similar words closer together in this space in order to maximize the predictive accuracy of their contexts.

One benefit of the continuous skip-gram model is that it creates representations which capture some local context as in the DepVec model, which is required to make predictions about the previous and next tokens. However, it must also encode some topical knowledge in order to make accurate predictions about the words seven tokens away. Therefore, using the latent term representations from the continuous skip-gram model as a vector space puts it in a convenient position in between the two others we presented.

4.4 Corpus Processing

The vector models described above were developed using web-scale corpora collected from a combination of frequently used NLP corpora and web crawls on news websites. Table 4 indicates the number of documents used for each language along with their source. These corpora were part-of-speech tagged with in-house POS taggers for English and Spanish, TreeTagger[3] for Russian, and hunpos[4] for Farsi. The open-source MaltParser was used to produce dependency parses for all four languages (Nivre, 2003). Dependency counts for all words occurring fewer than 40 times and for triples occurring fewer than three times were discarded to minimize noise.

5 Concept Induction and CM Recognition

In Section 4, we described our DepVec representation of terms as vectors in a high-dimensional distributional space. These vector representations encode both the dominant grammatical contexts of a term as well as the selectional preference information associated with it in the form of G-test scores. In this section, we describe our methodology for inducing conceptual domains for a linguistic metaphor by adapting techniques for unsupervised word-sense induction (Erk and Padó, 2008; Korkontzelos and Manandhar, 2010; Hope and Keller, 2013). In particular, we induce conceptual domains in an unconstrained manner by extracting the grammatical co-occurrents of an LM source term (i.e., the 'concept candidates') and clustering them into semantically-related concept clusters. Both the clusters and our

---

[3] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[4] http://code.google.com/p/hunpos/

1758

given source domains are then mapped into a distributional vector space, allowing us to compute cluster-to-domain scores. Finally, each source domain is assigned a score based on its affinity to each individual cluster with these affinity scores weighted according to cluster quality. This results in an overall weighted ranking of the given source conceptual domains for the linguistic metaphor.

5.1 Extracting Concept Candidates

Given a linguistic metaphor which consists of a metaphor source, s (e.g., "cure"), and a metaphor target, t (e.g., "poverty"), our system extracts a set of terms (i.e., "concept candidates") from the typical grammatical contexts of s as found in the web-scale corpus described in Section 4.4. In order to extract these candidates, we first determine the syntactic relation, r, which exists between s and t. This relation is the key point of interaction between the domains of the source and the target for the given LM and, as such, it provides an indication of which terms will contribute the most to our understanding of the underlying conceptual mapping. In addition, we make use of a predefined set of relations that are semantically meaningful – specifically the subjects and objects of verbs (i.e., "nsubj", "nsubjpass", and "dobj"),[5] attributes and verbs associated with nouns (i.e., "amod", "dobj^{-1}", "nsubj^{-1}", and "nsubjpass^{-1}"), the terms modified by adjectives or adverbs (i.e., "advmod^{-1}" and "amod^{-1}"), and prepositional relations (e.g., "prep_by", "prep_of", "prep_for"). Using this set of relations, R, we extract the set of candidate terms, X, that have been found to co-occur with the term s within some relation $r_i \in R$ in the preprocessed, web-scale corpus described in Section 4.4 such that $X = \{x | (s, r_i, x) \text{exists in the corpus}\}$.

To improve the quality of our extracted candidates, we apply three criteria to isolate those that best exemplify the underlying non-metaphorical senses of s. First, we anticipate that any term in X which does not co-occur with s at least k times will not be informative,[6] and so we remove such terms from further processing. Next, we predict that poorly imageable terms (i.e., highly abstract terms) are likely to represent metaphorical usages of s and so are unlikely to be integral to a given literal source domain, so these are filtered out as well.[7] Finally, to improve our ability to map these candidates into a conceptual domain, we remove terms that are not significantly related to any of our provided source domains (i.e., those that are off-topic) along with terms that are strongly related to multiple source domains (i.e., those that are ambiguous) as these provide little evidence to distinguish the most appropriate concept for the given LM.[8] We determine the relatedness of a term to a source domain by measuring the similarity of the term and domain vectors in our distributional space as described in Section 5.3.

5.2 Clustering Concept Candidates

Once the candidates have been extracted, they are clustered using a hierarchical agglomerative clustering algorithm with the distance metric defined as the cosine distance between the vectors within one of our distributional vector spaces. Each cluster is then assigned a quality score based on its size (to prefer large clusters with a large amount of semantic evidence), average internal distance (to prefer tighter clusters), and co-occurrence frequency with the LM source (to prefer more closely related terms). Formally, we define the weight associated with a given cluster using the following equation:

$$w(C) = (1 - IDIST(C)) * (S2(C) + FREQ(C) * (1 + S2(C)))$$

$$S2(C) = \frac{\max(SIZE(C)^2, k)}{k}$$

where $IDIST(C)$ represents the average vector distance between all pairs of terms in cluster C, $FREQ(C)$ represents the total co-occurrence frequency of the terms in C with the original LM,

---

[5]These dependency relation types come from the MaltParser.

[6]We empirically set k to 3.

[7]We estimate candidate imageability by combining the scores of the candidate's most distributionally similar words for which an imageability score is available in the MRC psycholinguistics database (Coltheart, 1981) using the ranked weighting methodology described in Mohler et al. (2014).

[8]Note that filtering by conceptual domain relatedness is only necessary when mapping the induced concepts to a predefined set of source concepts.

$SIZE(C)$ represents the number of unique terms in C, and k is a tuning parameter meant to favor large clusters.[9] Singleton clusters are discarded.

5.3 Assigning Domain Scores to Concept Candidates

We propose two methods for calculating domain scores for candidates – one which attempts to compare candidate vectors to a source domain directly, and and another which attempts to compare them to individual facets of the domain. These two methods rely on representing sources [CentS], or facets [CentF], as centroids which take the average of the vectors of each the lexemes assigned to that source (or facet). Our three vector spaces – DepVec, W2V, and LSA – along with our two methods for mapping terms to domains – CentS and CentF – correspond to six approaches to modeling a CM domain in some vector space.

In each case, the result for a given candidate is a distribution over all source domain scores. This distribution is then normalized by subtracting the mean score between the candidate vector and any of the source concepts. Formally, we define the normalized distribution for concept candidate x as:

$$S(x, D_y) = (1 - DIST(x, D_y)) - \frac{\sum\limits_{D_k \in D} (1 - DIST(x, D_k))}{|D|}$$

where D is defined as the set of all known source domains and $DIST(x, d)$ is the cosine distance from x to a CM domain d in one of our vector spaces.

5.3.1 Assigning Domain Scores to Clusters

Within a given cluster (found as described in Section 5.2), the individual concept domain scores can then be combined to produce cluster-level domain scores. For a given cluster C_x, the score associated with a particular source domain D_y is defined as follows:

$$S(C_x, D_y) = \sum_{i=1}^{N} \frac{S(C_{xi}, D_y)}{\alpha^i}$$

where N represents the number of concepts in C_x with a positive score for the domain D_y, C_{xi} is the i-th highest score associated with any candidate in the cluster, and α is a tuning parameter which bounds the growth of the cluster-level score.[10] Any cluster with a maximum domain score that does not exceed a threshold is discarded as being weakly related to any CM source domain.

5.3.2 Assigning Domain Scores to the Linguistic Metaphor

We then sum the cluster-level source domain scores, scaling each by its associated cluster quality weight $w(c)$ as computed in Section 5.2. By scaling cluster domain scores in this way, we ensure that the most pure and discriminating clusters contribute the most to the overall LM domain scores. The final result measuring the association between the given LM and the source domain D_y is then defined as:

$$S(D_y) = \sum_{C_x \in C} w(C_x) * S(C_x, D_y)$$

Applied across all known domains, we therefore produce a ranked and scored list of CM source domains (i.e., a mapping) that are associated with the given linguistic metaphor and can be used to drive more robust interpretation of the metaphor.

6 Evaluation

We evaluate two aspects of our end-to-end CM recognition system. First, we analyze the impact of our choice of vector space. Specifically, we compare the use of our DepVec space to link concept candidates

---

[9]In our experiments, k is set to 5.

[10]We have used a value of $\alpha = 2$ which ensures that the result remains within the bounds [0.0,1.0].

with source domains against two off-the-shelf vector space models – the continuous skip-gram model [W2V] (Mikolov et al., 2013) [11] and latent semantic analysis [LSA] (Landauer and Dumais, 1997). Both alternative models were trained over the same corpus as in our DepVec space using a predefined number of dimensions (300 for W2V; 400 for LSA). Second, we have experimented with two different metrics for calculating the distance between a vector and a source concept – the cosine distance to the source-level centroid (CentS) and the cosine distance to the facet-level centroid (CentF).

Our evaluation dataset consists of a held out, unseen set of documents taken from a variety of news articles, opinion pages, and blogs on the open web. These documents consist of 3 to 5 sentences each and cover four of our focus languages.[12] They were then annotated by two native-proficiency speakers in the following way. For each LM, they were instructed to choose the most closely related source concept from our list of 51 provided. Any source concepts selected by at least one annotator were considered correct. Since our CM recognition system produces a ranked list of source concepts, we report both the accuracy associated with our top-ranked concept and the accuracy of the system when allowed to select two.

| | | Cluster Linking | | | | | | | |
| | | English | | Spanish | | Russian | | Farsi | |
Vector Space	Distance	Acc@1	Acc@2	Acc@1	Acc@2	Acc@1	Acc@2	Acc@1	Acc@2
DepVec	CentS	28.0%	44.1%	33.3%	43.4%	24.4%	32.6%	16.5%	27.5%
	CentF	25.8%	40.9%	33.3%	49.4%	25.6%	34.9%	26.4%	40.7%
LSA	CentS	34.4%	45.2%	31.0%	41.4%	27.9%	41.9%	22.0%	27.5%
	CentF	**38.7%**	**54.9%**	27.6%	46.0%	29.1%	47.7%	31.9%	44.0%
W2V	CentS	24.7%	36.6%	42.5%	55.2%	31.4%	43.0%	25.3%	34.1%
	CentF	28.0%	44.1%	**46.0%**	**58.6%**	**34.9%**	**48.8%**	**35.2%**	**48.4%**

Table 5: The accuracy of our conceptual interpretation system. We experiment with three vector spaces (LSA, W2V, and DepVec) and two source concept centroid representations – source-level (CentS) and facet-level (CentF).

These results indicate that the continuous skip-gram vector space [W2V] is well suited to the task of cluster-level concept mapping, consistently and significantly outperforming both the LSA space and the DepVec space in every language but English. We believe that this is a result of its probabilistic representation of local context which implicitly collects many of the same relations as the DepVec model while incorporating the advantages associated with dimensionality reduction which has not been incorporated into our DepVec model.[13] We further observe an unmistakable dominance of the facet-level centroid representation over the source-level representation. Based on these results, we believe that we have successfully demonstrated the contribution of our system's vector-space clustering component which groups concept candidates at a facet-level granularity.

7 Conclusion

In this paper, we have presented a novel approach to the problem of multilingual conceptual metaphor recognition which combines facet-based concept induction with a distributional vector space representation of metaphor. We have experimentally demonstrated the advantage of our fine-grained concept induction approach within a variety of vector space models, including our novel DepVec space. Taken together, we hypothesize that a facet-level conceptual model represented in a relational context vector space will serve as a reliable foundation enabling high-quality metaphoric interpretation in future metaphor research. Future work includes expanding the set of concept candidates through higher-order dependency contexts, improved clustering techniques, and evaluating the induced clusters directly.

---

[11]We make use of the implementation included as part of the gensim python package: `http://radimrehurek.com/gensim/`

[12]This dataset consists of the following counts of documents: English (92), Spanish (86), Russian (85), Farsi (90).

[13]During our pilot experiments, we applied singular value decomposition (SVD) to the DepVec space without any significant improvement to system performance.

Acknowledgments

References

Pierre Beust, Stéphane Ferrari, Vincent Perlerin, et al. 2003. NLP model and tools for detecting and interpreting metaphors in domain-specific corpora. In Proceedings of the Corpus Linguistics 2003 conference, volume 16, pages 114–123. Citeseer.

Danushka Bollegala and Ekaterina Shutova. 2013. Metaphor interpretation using paraphrases extracted from the web. PloS one, 8(9):e74304.

D. Bracewell, M. Tomlinson, M. Mohler, and B. Rink. 2014. A tiered approach to the recognition of metaphor. In Computational Linguistics and Intelligent Text Processing.

Siaw-Fong Chung, Kathleen Ahrens, and Chu-Ren Huang. 2005. Source domains as concept domains in metaphorical expressions. International Journal of Computational Linguistics and Chinese Language Processing, 10(4):553–570.

Max Coltheart. 1981. The MRC psycholinguistic database. The Quarterly Journal of Experimental Psychology, 33(4):497–505.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 897–906. Association for Computational Linguistics.

J. Feldman and S. Narayanan. 2004. Embodied meaning in a neural theory of language. Brain and language, 89(2):385–392.

Lisa Gandy, Nadji Allan, Mark Atallah, Ophir Frieder, Newton Howard, Sergey Kanareykin, Moshe Koppel, Mark Last, Yair Neuman, and Shlomo Argamon. 2013. Automatic identification of conceptual metaphors with limited knowledge. In Twenty-Seventh AAAI Conference on Artificial Intelligence.

David Hope and Bill Keller. 2013. MaxMax: a graph-based soft clustering algorithm applied to word sense induction. In Computational Linguistics and Intelligent Text Processing, pages 368–381. Springer.

Ioannis Korkontzelos and Suresh Manandhar. 2010. UoY: Graphs of unambiguous vertices for word sense induction and disambiguation. In Proceedings of the 5th international workshop on semantic evaluation, pages 355–358. Association for Computational Linguistics.

G. Lakoff and M. Johnson. 1980. Metaphors we live by, volume 111. Chicago London.

G. Lakoff. 1993. The contemporary theory of metaphor. Metaphor and thought, 2:202–251.

T.K. Landauer and S.T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review; Psychological Review, 104(2):211.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In Proceedings of the 17th international conference on Computational linguistics-Volume 2, pages 768–774. Association for Computational Linguistics.

Z.J. Mason. 2004. CorMet: A computational, corpus-based conventional metaphor extraction system. Computational Linguistics, 30(1):23–44.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Michael Mohler, Marc Tomlinson, and David Bracewell. 2013. Applying textual entailment to the interpretation of metaphor. In IEEE Seventh International Conference on Semantic Computing (ICSC), pages 118–125. IEEE.

Michael Mohler, Marc Tomlinson, David Bracewell, and Bryan Rink. 2014. Semi-supervised methods for expanding psycholinguistics norms by integrating distributional similarity with the structure of WordNet. Language Resources and Evaluation Conference 2014.

Joakim Nivre. 2003. An efficient algorithm for projective dependency parsing. In Proceedings of the 8th International Workshop on Parsing Technologies (IWPT. Citeseer.

Astrid Reining and Birte Lönneker-Rodman. 2007. Corpus-driven metaphor harvesting. In Proceedings of the Workshop on Computational Approaches to Figurative Language, pages 5–12. Association for Computational Linguistics.

Hinrich Schütze. 1998. Automatic word sense discrimination. Computational linguistics, 24(1):97–123.

Ekaterina Shutova and Lin Sun. 2013. Unsupervised metaphor identification using hierarchical graph factorization clustering. In Proceedings of NAACL-HLT, pages 978–988.

E. Shutova, L. Sun, and A. Korhonen. 2010. Metaphor identification using verb and noun clustering. In Proceedings of the 23rd International Conference on Computational Linguistics, pages 1002–1010. Association for Computational Linguistics.

Ekaterina Shutova. 2010. Automatic metaphor interpretation as a paraphrasing task. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 1029–1037. Association for Computational Linguistics.

Tomek Strzalkowski, George Aaron Broadwell, Sarah Taylor, Laurie Feldman, Boris Yamrom, Samira Shaikh, Ting Liu, Kit Cho, Umit Boz, Ignacio Cases, et al. 2013. Robust extraction of metaphors from novel data. Meta4NLP 2013, page 67.

T. Veale and Y. Hao. 2008. A fluid knowledge representation for understanding and generating creative metaphors. In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, pages 945–952. Association for Computational Linguistics.

Yorick Wilks, Lucian Galescu, James Allen, and Adam Dalton. 2013. Automatic metaphor detection using large-scale lexical resources and conventional metaphor extraction. Meta4NLP 2013, page 36.