Mapping conceptual features to referential properties

Abhijeet Gupta*, Gemma Boleda†, Marco Baroni‡ and Sebastian Padó*
*IMS, University of Stuttgart
Email: {abhijeet.gupta,sebastian.pado}@ims.uni-stuttgart.de
†Universitat Pompeu Fabra
Email: gemma.boleda@upf.edu
‡CIMeC, University of Trento
Email: marco.baroni@unitn.it

***Abstract*—We report on initial work on bridging the concept-to-reference gap using distributional semantics. Specifically, we aim at predicting properties of countries, using distributional vectors to infer database information. Our results are highly encouraging, since we achieve an error reduction of 30% over the baseline and are not far from the upper bound.**

I. Introduction

Words, such as *dog*, encode concepts that can apply to many, diversely individuated entities (poodles, dalmatians, old dogs, young dogs, etc.). This is so because humans need to describe the daunting variety of concepts in the world with a limited shared vocabulary [1]. As a result, words are notoriously vague, that is, they do not have well defined boundaries regarding the objects that they apply to [2]. However, when used in a specific discourse, words and more complex linguistic expressions are often used to refer to specific entities with fixed properties, such as "this dog" when used for my neighbor's dog at 3.15pm in the afternoon on April 30th 2015 in Barcelona, a dog that is brown and quiet and not very large. Theoretical and computational models of meaning need to account both for conceptual and referential aspects of meaning, and current models are typically lacking on one or the other end.

For instance, distributional semantic models [3], [4] excel at modeling conceptual aspects: They predict, among many other things, that *dog* and *cat* are more similar than, say, *dog* and *stone*. They also predict generic properties of concepts, such as mammals being typically furry [5]. However, they have no handle on specific *instances* of a concept, on identifying a particular dog or describing its properties; that is, they skirt the question of reference – with [6] being a notable exception.

This abstract reports on initial work on bridging the concept-to-reference gap using distributional semantics. Specifically, we predict properties of countries, using distributional vectors to infer database information. Countries are individual entities denoted by proper nouns. Databases (we use FreeBase[1]) encode properties of entities: For countries, that includes their capital city, their geolocation, or their GDP. Note that properties can be either binary (for a given country and a given city, it either holds that the city is the country's capital, or not) or continuous (like a country's GDP). We use generic distributional information to predict specific properties of countries. We do so by learning a mapping from distributional vectors to FreeBase vectors, that is, a function that learns, given a distributional vector for a country, what its FreeBase feature values are. For the moment, we are using a simple logistic regression model, though we plan to train a neural network in the future. Our results are highly encouraging, since we achieve an error reduction of 30% over the baseline and are not far from the upper bound.

[1] http://www.freebase.com.

II. The Model

For our mapping experiments we use a Multivariate Logistic Regression (MLR) model. The model takes a given set of features as input and generates a predetermined set of outputs as predictions. Each input feature is converted to a weighted input to which a transformation function is applied to generate the predictions. The choice of transformation function for the MLR model depends on the type of the values given as input. Since our input space consists of binary as well as real-valued features, we have used the sigmoid or logistic function for the sake of consistency. This function transforms a value t onto:

$$S(t) = \frac{1}{1 + e^{-t}} \tag{1}$$

To estimate the weights, we use gradient descent. As error function, we use Cross Entropy (Equation 2), which can accurately measure small valued probabilities. In the Equation, N is the total number of training examples, y_n the gold value of a given feature for example n, and p_n the value predicted according to the current weight set. To prevent overfitting, we perform an L_2 regularization with a lambda parameter of 10^{-5}.

$$E = -\frac{1}{N}\sum_{1}^{N}[y_n \log p_n + (1 - y_n)\log(1 - p_n)] \tag{2}$$

We present two instantiations of the MLR model. The first model (called FB2FB), serving as an upper bound, takes Freebase features as input and predicts the same features as outputs. The second model (called Vec2FB), takes distributional vectors and predict the Freebase features as output. We estimate a lower bound through an informed baseline which computes predictions as follows: for binary features, we predict the majority class (0 or 1); for continuous features, we predict the mean value of the feature in the test set.

III. Setup and Evaluation

As distributional vectors for the countries, we use pre-trained 1000-dimensional vectors for FreeBase entities,[2] produced with the skip-gram version of Mikolov et al.'s neural language model on a 100-billion token news corpus [7].

Our country dataset consists of the 260 countries in FreeBase that have a distributional representation in the dataset (some countries do not exist anymore, like Yugoslavia, but we keep them in the dataset because they are still useful for our purposes). We use a standard 60%-20%-20% partition into training, development, and test set, although for the moment we do not optimize parameters on the development set.

We predict the 745 FreeBase features (125 binary, 620 continuous) that have a value for at least 30 countries. Examples of binary features are CONTAINEDBY::EUROPE and MEMBER-OF::ORGANIZATION::WORLD BANK, whereas examples of continuous features are GEOLOCATION::LONGITUDE and POPULATION::2011::NUMBER (meaning "population in 2011"; we have different estimates for different years).

Since there is no appropriate unified evaluation measure for the types of features that we use (binary and continuous), we evaluate them separately. For binary features, we report standard accuracy. For continuous features, we define the *normalized rank score (NRS)* as the average rank difference between gold and the predicted values, normalized by the length of the list. The intuition is that a predicted value of a feature for a country c, $p_f(c)$ should be "close" to the gold value $g_f(c)$, but that we care about closeness in rank more than about closeness in absolute value. As an example, take the POPULATION::2011::NUMBER feature, and imagine that we only have three countries (Germany: 80M; Spain: 36M; and Netherlands: 17M). If we predict 56M for Spain's population, it is still (correctly) predicted as the second most populous country (rank difference of 0); a prediction of 16M, however, would push Spain to third place (rank difference of 1).

Let Tr_f be the set of training examples for features f, and Ts_f be the set of test examples, and $rank(v, S)$ the rank of value v in set S. Then we define

$$NRS(f) = 1/||Ts_f|| \sum_{t \in Ts_f} |rank(p_f(t), Tr_f \cup Ts(f)) - rank(g_f(t), Tr_f \cup Ts(f))|/||Ts_f||$$

The normalization is necessary because not all features are instantiated for all countries. This yields a measure with range [0..1], smaller numbers indicating better ranking, that can be understood as a variant of Mean Reciprocal Rank (MRR). Note that we include all countries in the ranked list, no matter if they are in the training, development, or test set. This is to make it harder for the model to obtain good results, since we are adding more confounders.

IV. Results

Table I shows the results of the experiment. Recall that both evaluation measures range between 0 and 1; for accuracy 1 is best and for normalized rank 0 is best.

[2]Available at https://code.google.com/p/word2vec/.

TABLE I. Results

	Accuracy (binary features)	Normalized Rank Score (continuous features)
Baseline (lower bound)	0.79	0.34
Vec2FB	0.86	0.24
FB2FB (upper bound)	0.97	0.16

Recall that the upper bound model, FB2FB, is trained on FreeBase features and predicts FreeBase features. All it has to learn, therefore, is that there is one feature in the input that is the right value for the output. This yields an accuracy of 0.97 and a normalized rank of 0.16. The baseline is at 0.79 accuracy and 0.34 normalized rank. These are already fairly competitive values due to the informed nature of the baseline (cf. Section II).

The model whose performance we are actually interested in, Vec2FB, in which we map from distributional information to FreeBase features, achieves .86 accuracy and .24 normalized rank. Both represent around 30% error reduction over the baseline, and are not very far from the upper bound. We feel that this a promising result, given that the FreeBase features we predict are fairly fine-grained, and we only use generic distributional information as input.

V. Related work

There are essentially three lines of related research in the literature we know of.

a) Reference within distributional semantics.: Since the main topic of this abstract has been largely ignored so far, we only know of one piece of related work, Herbelot's paper at this year's IWCS [6]. In this paper, she uses vectors for the main category of the entities she tackles (for instance, "man" for Mr. Darcy, one of the characters in the novel *Pride and Prejudice*), and individualizes them by combining them with the vector for the entity (e.g., the "Mr. Darcy" vector as extracted from the novel) using word meaning in context methods. She observes that directly using the entity vectors does not work with her metrics. In contrast, we directly use distributional vectors for countries, and learn a mapping from distributional to FreeBase vectors.

b) Mapping between different spaces.: There has recently been quite a lot of work on mapping between different modalities, for instance from textual to visual data [8] or from textual to perceptual properties represented as feature norms [9]. We map from textual to database information with the goal of approximating reference.

c) Combining distributional semantics and databases.: The work of Freitas and colleagues [10], [11] also combines distributional information with databases. However, they do not learn a mapping and they use distributional information to guide database query searches. Socher et al. [12] is perhaps the closest work to ours, in that they exploit distributional vectors to learn about database properties, but their focus is on *relations* between entities, rather than on learning discrete and continuous features associated with a single entity.

VI. Conclusion

This abstract reports ongoing work on bridging the concept-to-reference gap with distributional semantics. We have shown that it is possible to use distributional information, which is based on contexts of use and provides generic information, to predict specific properties of entities as encoded in a database.

Our model is still very simple, and we have only tested it on one type of entity, namely countries. Our next steps will therefore be to use a different type of entity (cities) and a neural network instead of logistic regression. We are also analyzing the data (word2vec vs. FreeBase, as well as the results) to gain further insight into the task, the model, and the phenomenon.

More generally, this is just a first step in a larger research program. The interplay between the conceptual and referential modes of language is an area that requires more attention both from a linguistic and from a computational point of view.

References

[1] G. L. Murphy, *The Big Book of Concepts*. Cambridge, MA (etc.): The MIT Press, 2002.

[2] R. Keefe, *Vagueness*. Cambridge: Cambridge University Press, 2000.

[3] P. D. Turney and P. Pantel, "From Frequency to Meaning : Vector Space Models of Semantics," *Journal of Artificial Intelligence Research*, vol. 37, pp. 141–188, 2010.

[4] K. Erk, "Vector space models of word meaning and phrase meaning: a survey," *Language and Linguistics Compass*, vol. 6, no. 10, pp. 635–653, October 2012.

[5] M. Baroni, B. Murphy, E. Barbu, and M. Poesio, "Strudel: a corpus-based semantic model based on properties and types." *Cognitive science*, vol. 34, no. 2, pp. 222–54, Mar. 2010.

[6] A. Herbelot, "Mr Darcy and Mr Toad, gentlemen: distributional names and their kinds," in *Proceedings of the 11th International Conference on Computational Semantics*. London, UK: Association for Computational Linguistics, Apr. 2015, pp. 151–161.

[7] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic Regularities in Continuous Space Word Representations," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta: Association for Computational Linguistics, 2013, pp. 746–751.

[8] A. Lazaridou, E. Bruni, and M. Baroni, "Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, Jun. 2014, pp. 1403–1414. [Online]. Available: http://www.aclweb.org/anthology/P14-1132

[9] B. T. Johns and M. N. Jones, "Perceptual inference through global lexical similarity," *Topics in Cognitive Science*, vol. 4, no. 1, pp. 103–120, 2012. [Online]. Available: http://dx.doi.org/10.1111/j.1756-8765.2011.01176.x

[10] A. Freitas and E. Curry, "Natural language queries over heterogeneous linked data graphs: A distributional-compositional semantics approach," in *Proceedings of the 19th international conference on Intelligent User Interfaces*. ACM, 2014, pp. 279–288.

[11] A. Freitas, J. a. C. P. da Silva, E. Curry, and P. Buitelaar, "A Distributional Semantics Approach for Selective Reasoning on Commonsense Graph Knowledge Bases," in *Natural Language Processing and Information Systems*. Springer International Publishing, 2014, pp. 21–32.

[12] R. Socher, D. Chen, C. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledge base completion," in *Proceedings of NIPS*, Lake Tahoe, NV, 2013, pp. 926–934.