Roleo: visualising thematic fit spaces on the web

Asad Sayeed and **Xudong Hong** and **Vera Demberg**
Cluster of Excellence "Multimodal Computing and Interaction"
Saarland University
66123 Saarbrücken, Germany
{asayeed,xhong,vera}@coli.uni-saarland.de

Abstract

In this paper, we present Roleo, a web tool for visualizing the vector spaces generated by the evaluation of distributional memory (DM) models over thematic fit judgements. A thematic fit judgement is a rating of the selectional preference of a verb for an argument that fills a given thematic role. The DM approach to thematic fit judgements involves the construction of a sub-space in which a prototypical role-filler can be built for comparison to the noun being judged. We describe a publicly-accessible web tool that allows for querying and exploring these spaces as well as a technique for visualizing thematic fit sub-spaces efficiently for web use.

1 Introduction

We developed Roleo as a web platform in order to visualize and explore the vector spaces generated by the process of thematic fit evaluation in distributional models. We define thematic fit to be a measure of the extent to which the selectional preference of a verb given a thematic role is fulfilled by a particular noun. For example, we expect "knife" to strongly fit the instrument role of "cut", but "sword" much less so, and "hammer" hardly at all. Modeling thematic fit has applications in areas like cognitive modeling and incremental parsing. Various efforts have produced human judgements of thematic fit for different combinations of verbs, roles, and nouns (Padó, 2007; Greenberg et al., 2015), and there have been a number of recent efforts to build models that correlate closely with those judgement datasets.

The most successful of these have been the Distributional Memory (DM) models (Baroni and Lenci, 2010), which are unsupervised models that produce sparse, very high-dimensional vector spaces. Recently, word embedding models with smaller numbers of dimensions have been tested, although they have yet to reach the degree of correlation with human judgements that DM models have (Baroni et al., 2014). Nevertheless, in both cases, some notion of geometric distance or similarity is used to substitute for the concept of fit.

If a geometric measure is used as the operational conceptualization of thematic fit, then we should be able to subjectively assess the quality of the space through visualization in order to gain a grasp, for example, of how easily the space is partitionable or clusterable. This capability is useful in the iterative engineering of features or for assessing the quality of training data.

A number of existing packages across many different development environments support low-dimensional projection and visualization of high-dimensional vector spaces. There are also a small number of web sites that allow word embeddings to be visualized in a low-dimensional space (Faruqui and Dyer, 2014). However, the best-performing work in vector-space thematic fit evaluation projects sub-spaces from a full tensor space given a verb and a role. Roleo is designed to query and visualize these sub-spaces in a manner that reflects the evaluation process.

Roleo is live and available for use at `http://roleo.coli.uni-saarland.de/` with two example models and an efficient visualization technique. We have furthermore made the code for it open source and available at `https://github.com/tony-hong/roleo`.

1.1 Design goals

Our goals for the Roleo software are to:

- Provide a web-based platform for the exploration of thematic fit sub-spaces based on dif-

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics—System Demonstrations, pages 139–144,
Berlin, Germany, August 7-12, 2016.

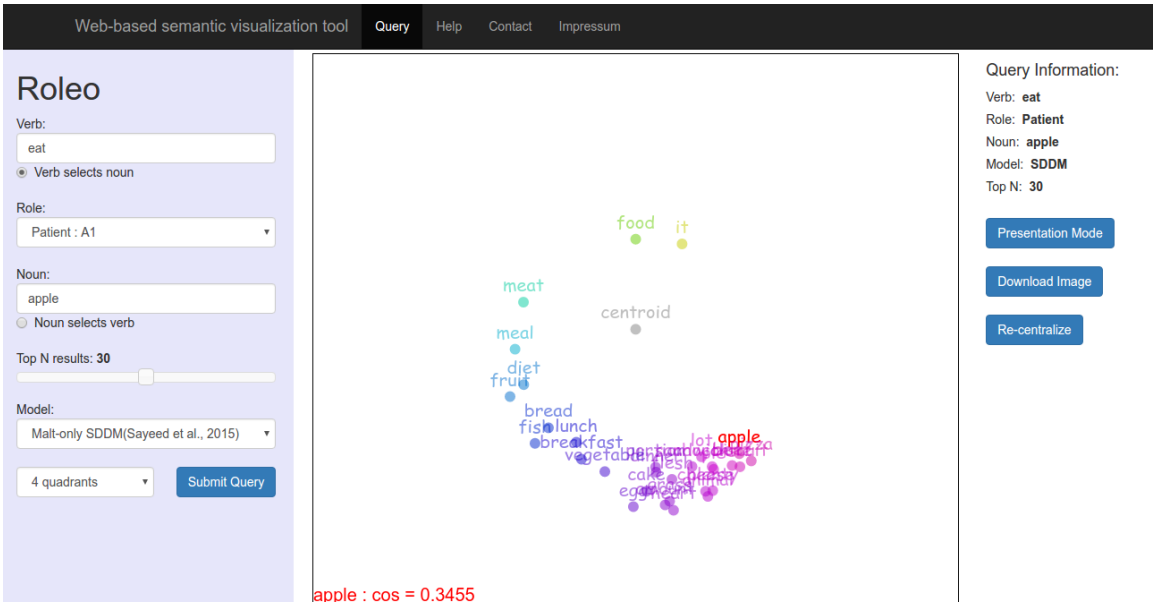

Figure 1: Initial screen on loading Roleo in a browser.

ferent vector-space modeling techniques. We begin with DM models.

- Make this type of semantic modeling accessible to other researchers via the web. This means that the interface must be reasonably user-friendly and allow visitors to test simple queries without knowing how to set all possible parameters.
- Facilitate presentations and demonstrations about thematic fit evaluation.
- Serve queries reasonably quickly, ideally at "web speed", so that it is reasonable to "play around" with the models. This puts a constraint on the kinds of projections and dimensionality reduction we can use.

2 Vector-space thematic fit modeling

2.1 Distributional Memory

The currently best-performing models on the thematic fit task, in terms of correlation with human judgements, are the Distributional Memory (DM) models, based on a technique first proposed by Baroni and Lenci (2010). A DM model is an order-3 tensor with two axes that represent words in the model's vocabulary and one axis that represents links between the words, so that the cell of the tensor is a positive real value for the occurrence of a triple <word0,link,word1>. That occurrence is an adjusted count of frequency such as Local Mutual Information (LMI). The link between the words is a connection acquired from processing a corpus, such as with a dependency parser.

This structure was extensively tested by Baroni and Lenci on a number of semantic tasks, including on thematic fit modeling. Their procedure for thematic fit modeling was the following: given a verb v and a role r, they look up all the nouns n such that each $< v, r, n >$ LMI is within the highest 20 for v and r. Then for each n, they get a word vector w_n from the model by looking up all the word0 and link contexts that n appeared in as word1; the vector is assembled from the LMI values in those cells of the tensor. Given a candidate noun m against which to evaluate the fit with v and r, a w_m vector is similarly found. All the <word0,link> contexts are the dimensions of a subspace of the space represented by the DM tensor as a whole; they can number potentially in the millions. All the w_n vectors are summed to form a centroid that represents a "prototype" noun for that verb and role. The thematic fit of m is evaluated via the cosine similarity of w_m and the centroid.

2.2 Provided models

In our demonstration version of Roleo, we provide two models, the TypeDM model from Baroni and Lenci and the "Malt-only SDDM" model from Sayeed et al. (2015). TypeDM is trained on multiple corpora (BNC, ukWaC, and Wikipedia) that have been downloaded and parsed by MaltParser. The links between words that are used to form TypeDM's link axis are derived from short

MaltParser dependency paths via a partly handcrafted rule set. As TypeDM's links are derived from a syntactic parser, we must simulate semantic roles by interpreting these links. Roleo allows for the query of agent roles (via subject links), patient roles (via object links), instrument roles (via the preposition "with"), and location roles (via the prepositions "in", "at", and "on").

Malt-only SDDM (just SDDM from now on) is derived from a set of corpora similar to that of TypeDM: BNC and ukWaC. The main difference between TypeDM and SDDM are the link types, which in SDDM are PropBank roles, derived from applying the SENNA semantic role labeller to the corpora. The links are therefore the PropBank roles that connect verbs to nouns. SENNA (Collobert and Weston, 2007), however, labels entire noun chunks with roles, often including adjectives and whole relative clauses. Sayeed et al. experiment with a number of algorithms for extracting the noun head or bare noun phrase; the best performing SENNA-based technique is to use the MaltParser dependencies produced by Baroni and Lenci, but simply as a guide for head-identification. Sayeed et al. show that PropBank-based roles and TypeDM roles help cover different aspects of the thematic fit problem.

This process can be trivially reversed to represent the plausible verbs given a noun-role combination and to produce a visualization thereof. We provide this functionality inside Roleo, although it has never so far been evaluated on any task.

3 Efficient projection

One of our design goals was to build a query tool that delivered results in times reasonable for the web with limited resources, i.e., a single-PC web server with a modern CPU. Because we are visualizing thematic fit sub-spaces constructed around a centroid, we also looked for a projection that puts the centroid at the center of the display consisting of the prototype nouns that were used to construct that centroid. We experimented with principal component analysis (PCA) and t-SNE (Van der Maaten and Hinton, 2008) and found that at DM-scale dimensionality, these took too long and were too computationally intensive to resolve a query in web-appropriate time.

For this reason, we came up with a two-dimensional projection specialized to our problem: Fraction-Cosine Vector Mapping (FCVM). This projection is easy to calculate directly from the support s_w of each word vector w in the DM model (LMI) given the role and the verb, the support of the centroid s_c (which is just the sum of the supports of all the vectors in the top n words for that verb-role combination), and the cosine c_w of the angle between the centroid and the vector.

Let V be the set of n highest supported word vectors for the given verb-role combination. Then for each vector, we can calculate its x and y coordinates in FCVM with the following procedure. The x coordinate for a projected word vector w is the sum of proportions of contributions to s_c of all words w' with a support $s_{w'} > s_w$, meaning that the more LMI-associated w is with the verb-role combination, the closer it is to the centroid along the x-axis. That is,

$$x_w = \sum_{w' \in V} \frac{s_{w'}}{s_c} \quad (1)$$

The y_w-coordinate is simply $1 - c_w$. This means that x_w and y_w are both in the interval $[0, 1]$ and sit in the upper right quadrant of the Cartesian plane, with the origin (corner) as the centroid. We convert these to polar coordinates (r_w, θ_w) and then optionally apply an adjustment to spread the points out. This adjustment is to multiply θ_w by a multiple of 4, sweeping them across Cartesian quadrants, in order to bring the centroid closer to a circular cloud of points representing word vectors. The factors in Roleo are 1, 4, and 32, with 4 as the default. The higher the factor, the more circular the cloud. We finally convert the polar coordinates back to rectangular. We also include an option to display the polar coordinates by directly interpreting them as rectangular coordinates. The points are also given a colour that is dependent on their θ after the multiplication factor is applied.

4 System implementation

Roleo was developed in Python using the Django web development package. The DMs are implemented as Pandas dataframes stored in indexed HDF5 tables for efficient lookup. Vector algebra is implemented in NumPy. The two-dimensional coordinates for the points that appear in the visualization are calculated server-side, currently implemented on our own host, which is a single recent PC. The image is drawn client-side and requires a recent browser (we test with Firefox and Chrome).

Queries to Roleo take 2-10 seconds, depending

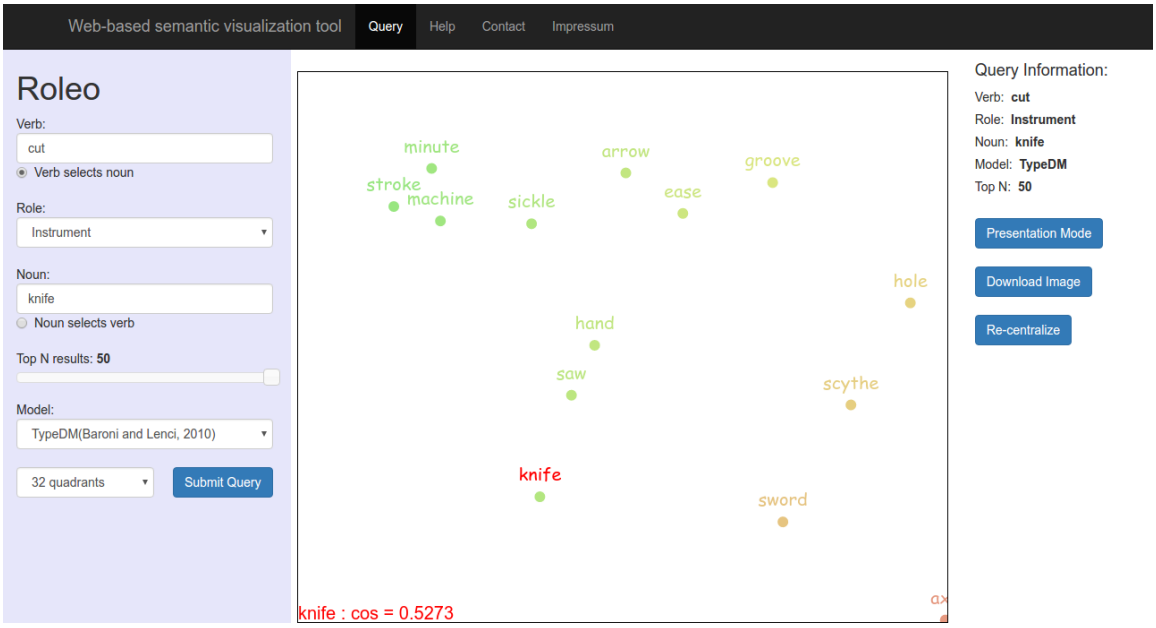

Figure 2: Zoomed-in query result for knife as instrument of cut under TypeDM with a 32-quadrant sweep and a space constructed from 50 prototype noun vectors.

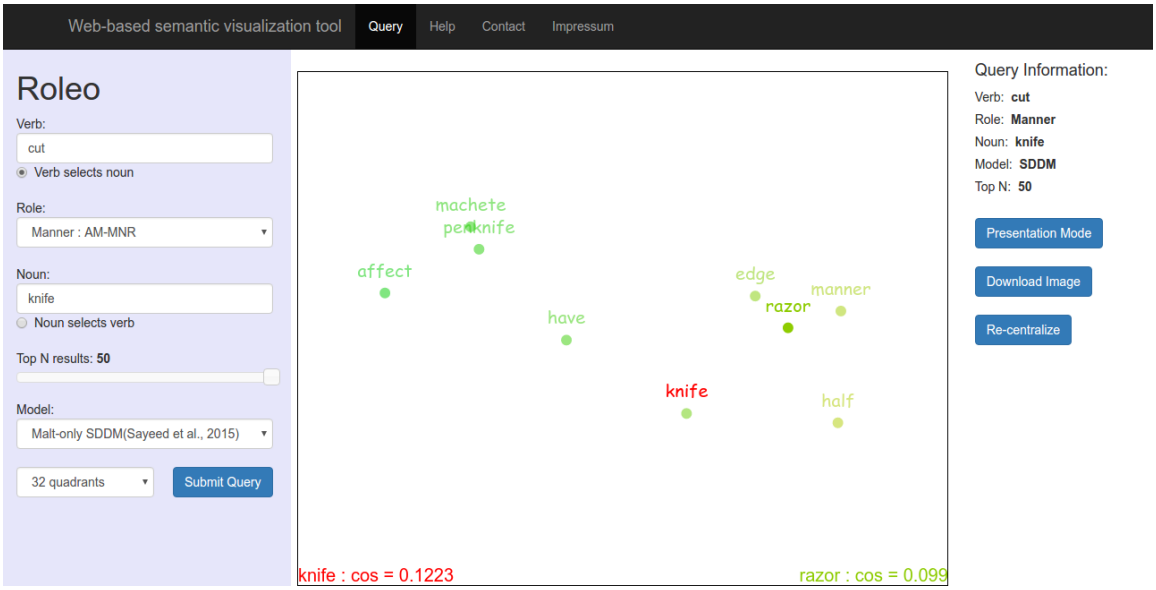

Figure 3: Zoomed-in query result for knife as ARGM-MNR of cut under SDDM with a 32-quadrant sweep and a space constructed from 50 prototype noun vectors. A touch gesture has highlighted the "razor" vector and put its cosine score on the bottom right corner.

on the number of vectors chosen by the user to form the centroid, within a tolerable range for a specialized web application.

4.1 Using Roleo

Roleo's initial screen on loading it for the first time in a browser is in figure 1. The screen is already populated with a query: how well "apple" fits as the patient of "eat"under SDDM, using 30 prototype nouns to calculate the centroid and populate the space. The 4-quadrant sweep is used to draw the canvas. Roleo is intended for use on a desktop PC or laptop or on a tablet.

Left pane Roleo's main options are shown on the left pane of the web page. There, the user can set the parameters and start the query. Fields are

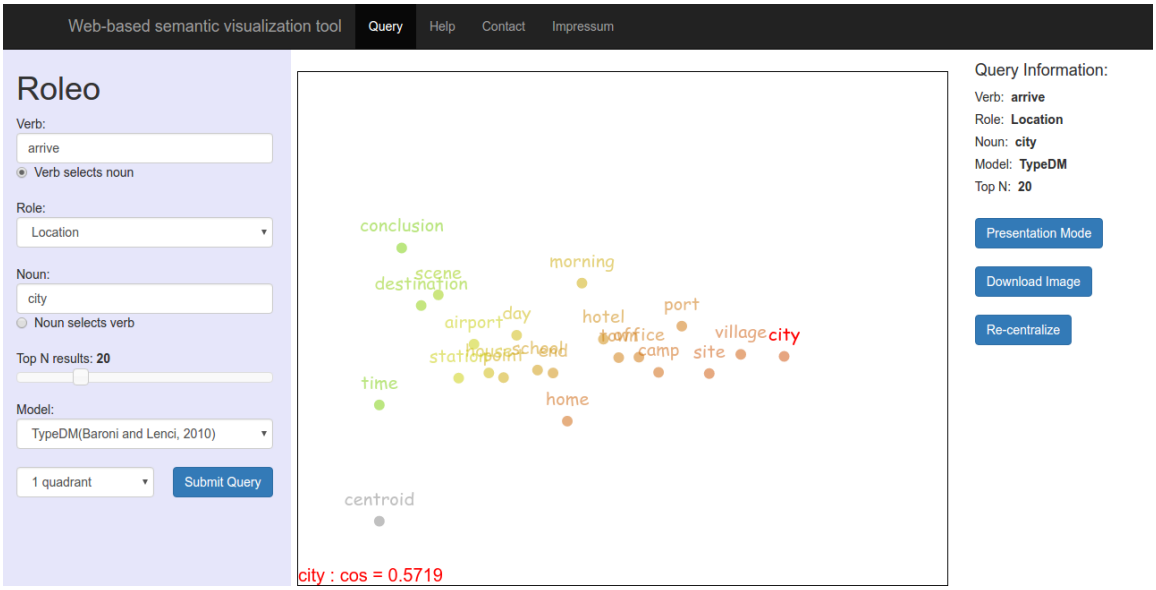

Figure 4: Query result (without zooming) for "city" as location of "arrive" under TypeDM with a 1-quadrant sweep and a space constructed from 20 prototype noun vectors.

available to enter a noun, a verb, and a role. The roles available are dependent on the model chosen. A slider allows the choice of between 10 and 50 top prototype vectors in increments of 10, and the choice of quadrant sweep size is available, including a "4-span by cosine" option, which is the direct interpretation of the polar coordinates.

The radio buttons "Verb selects noun" and "Noun selects verb" allows the user to set the direction in which the thematic fit query is executed. "Verb selects noun" is the default algorithm that chooses prototype nouns based on a verb-role combination. "Noun selects verb" allows the user to explore the choice of verbs based on a noun-role combination.

Main canvas The central pane of the Roleo page is the canvas on which the vectors are visualized. This pane can be scrolled and zoomed via mouse or touch gestures, depending on the user's browser, operating system, and hardware. The vectors are shown as small labeled circles on the canvas, with the gray dot as the centroid, usually located in the center for the 4- and 32-quadrant sweep displays. The queried vector is highlighted in red; the labels for the other vectors appear when the canvas decides there is space for them or when they are moused over. The bottom left corner contains the cosine similarity score (with the centroid) for the queried vector, and the bottom right corner displays the cosine similarity of a moused-over or touched vector.

Right pane Roleo's right pane contains the details of the query currently represented on the canvas, in case the user needs a reminder of the previous field contents as they change the fields to explore the space. In addition, it contains a button to shift Roleo into a full-screen presentation mode, a button to download the depicted space as an image file, and a button ("Re-centralize") to return the current query to its default view and cancel the effect of scrolling or zooming.

4.2 Example lookups

Figures 2 and 3 contain queries about "cut" and "knife" for TypeDM and SDDM respectively. For TypeDM, we chose the instrument role; for SDDM, we chose manner (PropBank "ARGM-MNR"). With TypeDM, we see items that are knife-like. Most of what appears there that is *not* knife-like can be used with the preposition "with", as we have defined the instrument role for querying TypeDM (section 2.2). Other parts of the space not depicted here contain less knife-like items, such as a region where "chainsaw", "clipper", "mower", and "grinder" are close to one another.

For SDDM, we also see knife-like instruments, but we see "half" and "manner", as in "cut in half" and "cut in a manner", also a result of PropBank. There is also probable noise in both cases (e.g.

"have", "hole"), as these spaces are ultimately derived from large corpora.

Figure 4 is a 1-quadrant, 20-prototype view of the "arrive"-location combination given a queried noun of "city" under TypeDM. This is therefore a rectangular view. Given the FCVM projection, "city" is in the middle of the group along the y-axis, meaning that it is the middle of the group for cosines. However, it is far along the x-axis, meaning that it had comparatively low LMI score with respect to "arrive" and the location role.

5 Demonstration

The centrepiece of our demonstration at the conference is a laptop or other computer display that allows conference visitors to interact with Roleo, as we explain its capabilities and advantages and explore different vector spaces with the help of an associated poster.

6 Future work

Roleo is under active development, and we intend to include significant additional features. Among these:

Adding models We plan to add more models, including newer, dense word-embedding spaces for comparison, in order to help us diagnose why these spaces seem to perform less well than DMs on the thematic fit task (Baroni et al., 2014).

More visualizations FCVM provides a way to project high-dimensional vector spaces down to two dimensions in a reasonable time for web use on a single thread on a single server. It principally represents the location of a vector with respect to the centroid, which is ideal for thematic fit modeling, and it leads to a tendency for vectors related via the verb to be close to one another. However, it loses a direct geometric or probabilistic interpretation of the proximity of vectors. Therefore, we are investigating the possibility of adapting FCVM and more processor-intensive procedures like t-SNE and PCA to one another. Currently, we are testing a lightweight SVD-based visualization algorithm that still centers points around the centroid; although it is more computationally intensive than FCVM, our preliminary observations are that it produces more well-defined clusters in acceptable time.

References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. volume 1, pages 238–247.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics* 36(4):673–721.

Ronan Collobert and Jason Weston. 2007. Fast semantic extraction using a novel neural network architecture. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, pages 560–567.

Manaal Faruqui and Chris Dyer. 2014. Community evaluation and exchange of word vectors at wordvectors.org. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Baltimore, USA.

Clayton Greenberg, Vera Demberg, and Asad Sayeed. 2015. Verb polysemy and frequency effects in thematic fit modeling. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*. Association for Computational Linguistics, Denver, Colorado, pages 48–57.

Ulrike Padó. 2007. *The integration of syntax and semantic plausibility in a wide-coverage model of human sentence processing*. Ph.D. thesis, Saarland University.

Asad Sayeed, Vera Demberg, and Pavel Shkadzko. 2015. An exploration of semantic features in an unsupervised thematic fit evaluation framework. In *IJCoL vol. 1, n. 1 december 2015: Emerging Topics at the First Italian Conference on Computational Linguistics*. Accademia University Press, pages 25–40.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research* 9(2579-2605):85.