Singular value decomposition (SVD) for MRS isomorphism

Glenn Slayden
University of Washington
gslayden@uw.edu

Abstract

I describe a new method for the automatic alignment of two or more Minimal Recursion Semantics (MRS) structures and deterministically scoring this alignment. The novel contribution is a technique for representing MRS structures in a matrix suitable for analysis via singular value decomposition (SVD). The SVD computation determines the optimal least-squares rotation—in a high-dimensional space—of the MRS structures relative to each other. Thus reduced, conventional vector space model analysis of MRSes is enabled. The general task of aligning arbitrary MRSes has important applications in machine translation, treebank bootstrapping for low-resource languages, and evaluating semantic mappings.

1 Introduction

Manipulating structures in the format of Minimal Recursion Semantics (MRS, Copestake et al. 2005) is a critical task in state-of-the art grammar engineering and computational semantics. In addition to serving as the cornerstone semantic formalism for the DELPH-IN consortium, MRS has also been adapted in non-HPSG environments, such as CCG.

This paper details ongoing work on a long-term project for developing bi-directional Thai-English machine translation using a system of semantic transfer between HPSG-based computational grammars. The current research hoped to demonstrate the transfer prototype. However, one stringent requirement was that this research include quantitative evaluation. Such measurements are non-trivial, and there are no unsupervised methods for measuring MRS similarity.

This paper is structured as follows. First I describe the test corpus, originally intended for research in bilingual semantic transfer, but also useful for illustrating a new, automatic method for the deterministic alignment of arbitrary, disjoint MRSes. The new method is described in the following section. Next, the transfer problem is addressed. I detail some test procedures in Section 5, and then use the new alignment technique is used to confirm the efficacy of a few simple improvements.

.

2 Test corpus and tools

To test semantic mapping and MRS isomorphism, a corpus of MRS pairings was developed. A set of 187 Thai sentences, each with a high-quality, human-authored English translation, was extracted from the database of the *thai-language.com* website, according to the criteria of obtaining one or more derivations when parsed by both of the grammars described below. The bitext sentence pairs which were parsed to obtain the semantic corpus are listed at the following link:

http://www.computational-semantics.com/ling575/sents.txt

DELPH-IN grammars and tools were used to parse the sentences. For the English sentence, MRS representations were produced by the English Resource Grammar (ERG, Flickinger 2000).[1] A small grammar of Thai, originally produced with the Grammar Matrix toolkit (Bender et al. 2002), and extended to handle a few additional linguistic phenomena, was used for Thai. The 187 simple sentences in the test corpus exercise the full competence of this small grammar. Naturally, the paired English translations of these sentences are not challenging for the ERG. The entire semantic corpus can be viewed at the following link. Also provided is a complete set of machine-readable text files with details on all the MRSes studied. Each 'mrs-dump' file contains data for *all* of the (one or more) MRSes generated from a single Thai or English sentence.

http://www.computational-semantics.com/ling575/index.html
http://www.computational-semantics.com/ling575/mrs-corpus

Note the flattened format of the MRS-corpus files, where variables which host properties seem to acquire the status of roles. The format was designed this way as a prelude to processing the MRS via the SVD method described in Section 4, but I found the compact and consistent form to be otherwise quite convenient. It also facilitates 'diff-style' comparison of MRSes somewhat, although typically the so-called *bag* nature of the primary MRS elements (namely, that they are in formally unordered sets) renders this practice futile.

DELPH-IN generally advocates native-orthography predicate naming, but to simplify this initial research, a ver-

[1] For both grammars, the Sem-I ('semantic interface') and VPM ('variable property mapping') modules were activated. These post-processors allow the grammar author to fine-tune and control the structure of the information published to MRS.

sion of the Thai grammar which uses English-language semantic predicate names was used. Minor bugs and issues identified in the Thai grammar during corpus preparation work were repaired (most notably in the VPM grammar, which was revamped).

The *agree* grammar engineering system (Slayden 2010) was used because it is able to load both grammars in the same process and coordinate their activities, simplifying experimentation.

3 Parse Selection

When first addressing the semantic corpus described in the previous section, the most immediate and pressing problem is that, unlike with the ERG, which has a well-developed stochastic parse-selection model, there is no such model available for Thai. As shown at the link above, there can be multiple readings for the English or Thai sentence, or both.

At first I enabled the English MaxEnt model and worked only with the top-ranked English sentence, but soon I discovered that often the reading with the highest score is not the desired one, or it represents a meaning which lies outside the competency of the Thai grammar. This type of problem was judged to be not in the scope of this project because it is not the job of a semantic mapping to repair intended variation.

Accordingly, I restored the exhaustive set of derivations for both Thai and English, and proceeded to seek a method for automatically forming a single, best, translation pairing from all of the available readings. This led to the successful development of the SVD technique, the main contribution of this paper, which is described in the next section.

4 Best alignment of arbitrary MRSes

In the process of seeking a sound evaluation paradigm for use in an extended study of declarative semantic transfer, I found that no automatic methods were mentioned in the MRS literature. Methods for gathering atomic 'triples' from MRSes are documented (Driden and Oepen, 2011), and these lists can be compared, but human interaction is required to set the expected order between the lists.

Formally, the elements in an MRS are taken as unordered, and this fact makes it difficult to establish the formal rigor of unsupervised methods. For the Thai-English translation case, this follows intuition as well, since asserting the correspondence of semantic entities which originate from different grammars seems perilous. Of the MRS isomorphism that I am aware of, none makes a claim of formal rigor. Ultimately, establish correspondences between elements requires manual assistance.

The problem dampens productivity in many areas. Stochastic training requires a large number of training instances of course, so the lack of a method for the fully-automatic alignment of MRSes is a major obstacle for bootstrapping with existing semantic resources. Consider the Redwoods Treebank, a large collection of annotated syntactic derivations for the ERG. Given accurate translation pairs, the method of unsupervised MRS selection shown here can isolate the intended semantic reading—as produced by the low-resource grammar itself—by checking semantic isomorphism against the top-ranked ERG semantics for the surface translation.

4.1 Singular Value Decomposition

In this section I describe a method for organizing the structure of an MRS within a two-dimensional matrix so that the mathematical technique of Singular Value Decomposition (SVD) can be applied. Intuitively, by examining co-occurring patterns in the input, the SVD reduction of a matrix arranges its rows and columns such that 'denser' portions are relatively distant from each other. The matrix form of the MRS is non-lossy, so the SVD analysis can capitalize on all its entity co-occurrences, and can model arbitrary substructure, coreferencing, and synthetic properties (e.g., VPM).

4.2 SVD Theory

SVD is two-mode factor analysis, allowing it to manipulate an $m \times n$ content matrix. By representing column vectors in a space where intersecting transitive co-occurrence relations constitute a cline of choosable dimensionality, the SVD simultaneously provides noise attenuation (smoothing), redundancy detection, and a similarity retrieval metric (Kontostathis and Pottenger, 2002). The singular value decomposition is defined as

$$A_{m\times n} = U_{m\times d}\Sigma_{d\times d}(V_{n\times d})^T$$

where $d = \min(m, n)$. The power of the decomposition exists in the fact that, as shown in Golub and van Loan (1996, 72-73), for any k, the maximum-likelihood 2-norm (least-squares) approximation of rank-deficient $A_{m\times k}$ is given by

$$\hat{A}_{m\times k} = U_{m\times k}\Sigma_{k\times k}(V_{n\times k})^T$$

This guarantee means that the k-dimensional space is usefully rotated so as to align the k axes in the k directions of greatest variation. The value for k is chosen empirically.

4.3 MRS adaptation

The challenge in adapting the SVD technique to the task of finding an optimal alignment between arbitrary graphs is in determining how to encode the graph in the two-dimensional matrix in a manner that preserves the transitive co-occurrence chains manifested by graph reentrancies. In the case of MRS, these co-occurrences are the variables, equivalence classes which freely span across MRS relations, establishing the semantic structure. Conventionally, the lossless representation of a directed graph seems to require a structure of dimensionality greater than 2.

The solution lies in recognizing that SVD's ability to evaluate transitive co-occurrence can be used to actually model the portion of the original MRS structure which exceeds the modeling power of a two-dimensional matrix—namely, the graph-reentrancies. Individual columns in the SVD typically represent distinct, unrelated instances—for example, the "documents" in a term-by-

document application. To model more complex structures, however, multiple columns can be designated to a single instance, and then logically joined by encoding a specific aspect of their internal structure in specially-designated rows. This is the approach to encoding MRS for SVD. Rather than assigning one SVD column to each relation, each relation spans several, and special SVD rows are dedicated to tying together the bundle. The novel idea is that the special role rows allow the important *in*-relation co-occurrence information to be preserved in the MRS-SVD encoding.

4.4 Layout

The details of the encoding is as follows. Tuples of $\langle MRS, relation, role \rangle$ form the columns. Thus a single relation (from either MRS) is spread across multiple SVD columns, one for each of its role positions (LBL, ARG0, RSTR, etc.).[2]

One special row is also dedicated to each role in the union of all roles present in the two inputs. The SVD array is marked with '1' where these intersect. After these special rows, there is one row for each variable, this time a union across both MRSes, *without* conflation. Singleton variable occurrences do not affect the calculation and need not be included. Array positions where variables occur are marked in the obvious way. An example MRS-SVD layout is shown at the following link. The conventional MRS notation for the two MRSes which are installed in the array is shown at the bottom of the image.

http://www.computational-semantics.com/ling575/mrs-svd.png

Recall that $\langle MRS, relation, role \rangle$ tuples form the columns of the SVD input matrix. In accordance with SVD convention as used in other NLP tasks, we compute only the 'right singular vectors,' V^T, and the columns designate our entities of primary interest. Therefore, it is for these tuples, and not MRS relations, that we will obtain the least-squares alignment, so the alignment will be computed without regard for the relations to which the $\langle relation, role \rangle$ tuples belong.

For the purposes of evaluating MRS isomorphism in the current research, this is not an issue, since the structure implied by the SVD result is never destined to be interpreted as a valid MRS. However, several schemes (such as clustering) for deriving a best *relation* alignment from the present result can be imagined. None were investigated, so these are left for future research.

4.5 Interpretation

Recalling the SVD definition above. In this application, m is the number of the distinct top-level roles across both MRSes *plus* the number of distinct, non-singleton variables across both MRSes, and n is number of relation-role tuples across both MRSes. According to this configuration, the left singular vectors U indicate the coordinates of each variable in the reduced k-space, the right singular vectors V give the coordinates of each relation (Σ is formed by arranging the singular values–which are the lengths of the principal semiaxes of the hyperelliptic projection of A (Golub and Loan, 1996, 71)—along the main diagonal of an otherwise empty $d \times d$ matrix).

Although results are still pending at the deadline for this report, it seems likely that one can take the value of the first singular value as an overall measure of the quality of the alignment.

An interesting property of SVD is that the (very) expensive reduction step can be performed prior to committing to the desired dimensionality; the result matrices simultaneously encode the optimal rotations for all k. The output matrices for a particular k value is trivially computed. At this point, one proceeds as if the two MRS were a high-dimensional vector space model (VSM)—with the caveat that the dimensions should not be considered interpretable (Schütze 1992). It is a simple matter to pair up columns according to closest cosine (vector) distance, remembering which columns originated from which $\langle MRS, relation, role \rangle$ tuple. This is the basic procedure used to evaluate baseline statistics for the Thai-English corpus. The details are given in the next section.

5 Test methodology

The SVD method of MRS alignment allows the alignment baseline of the raw corpus to be computed. These figures are shown. As noted in Section 4.4, the SVD provides the optimal set of tuple parings between each $\langle MRS, relation, role \rangle$ tuple in the source MRS and the target MRS. Important aspects of the MRS formalism—notably, for example, and as mentioned above, the binding of a role to its relation—are not respected during this process. Also not enforced in the current implementation is that a target $\langle relation, role \rangle$ tuple not be aligned with more than one source element. Effecting these requirements, and others, can be achieved with suitable adjustments to the VSM analysis stage which follows the SVD computation.

5.1 Evaluating role names and atomic types

Given an alignment produced by the SVD technique described above, it is trivial to evaluate precision and recall for semantic role[3] names, as well as constant or 'atomic' types. They do not interact with other parts of the MRS and can be counted in the obvious way.[4] In the current implementation, these results make no adjustment for the fact that more than one source tuple may be assigned to the same target tuple.

[2] The ordering of the rows and columns in the SVD matrix is irrelevant. Indeed, it is the task of the SVD to determine the optimal k-dimensional rotation of the input instances.

[3] Arbitrary TFS-style (typed feature structure) substructure below MRS semantic variables can be processed with the SVD scheme, so my use of "roles" should be understood to include the *features* in such "feature/value" pairs as well.

[4] I believe the MRS formalism forbids (e.g. TFS-effected) co-referencing between non-variable nodes, and the SVD layout plan outlined here assumes this.

5.2 Variables

If one were concerned with only the 'sub-type' of the semantic variables (e.g. `h`, `x`, `u`, `i`...), then, just like the role names and atomic types, evaluating the precision and recall of the alignment would be trivial. In fact, I report variable sub-type accuracy figures.[5] Unfortunately, however, even given an exhaustive node alignment from the source MRS to the target MRS such as the SVD technique provides, it is challenging to evaluate the success of mapping the variables. This is not simply because the variable numeric indices may be assigned differently in two logically identical structures, which is relatively simple to address. Rather, it is because the mapping introduces diverse cases for which interpretative decisions must be made.

For example, in the current demonstration, the source MRS will get full credit for a variable of the correct type that it places in the target *when none of the source's other alignments* contain that variable in the target. However, if the target has unmapped positions, now how should a repetition of that variable (that is invisible to the source) now be handled?

Another rare case is where the target has fewer variables than the source (and in fact, fewer $\langle relation, role \rangle$ tuples, which necessarily results in some target nodes being targeted by multiple source nodes). In this case, the source can (falsely?) be marked correct by re-using a target variable—if it happens to have the correct signature for multiple source variables.

In the current design, variables are also evaluated globally over an MRS, as opposed to by $\langle relation, role \rangle$ tuple. This means each variable in the MRS counts as only one opportunity to get all of its alignments (specifically, those that are visible to the source) correct. In this case, "correct" means that the target must the same variable at all the positions, *as mapped through the alignment*, where the source MRS does. (This more subtle analysis also requires that the variable 'sub-type' be correct at all of the mapped positions) Thus, "variable precision" is the number of wholly-correct variable signatures relative to the number of source variables, whereas "variable recall" is relative to the number of target variables.

5.3 Transfer

Work on transfer is still ongoing, in particular depending upon the completion of the unsupervised cross-language MRS alignment work which is described here. Fortunately, as demonstrated in this report, that work is essentially complete and has surpassed all expectations for its success.

6 Evaluation and results

Prior to using the SVD to identity the most isomorphic MRS from a list of candidates, I evaluated precision and recall for only and all of the translation pairs which generated—for both languages—just one derivation. This was the case for 39 of the 187 instances. A raw dump of the output from this run can be found here:

http://www.computational-semantics.com/ling575/1th-1en.html

This file, though perhaps opaque, summarizes the bulk of my work on this project. In each delimited section, an independent alignment is computed for both the Thai-to-English and English-to-Thai directions, with the $\langle relation, role \rangle$ match-ups shown in detail. Because the SVD is computationally expensive but the resulting VSM supports the extraction of diverse work products, it makes sense to compute both mappings at the same time.

The high-dimensional Euclidian distance between the source and target—according to the alignment—is also given, but this has not proved useful yet. More predictive is the first singular value computed by the SVD. It seems to be a measure of the alignment quality. The minimum value of 2.54832 occurs with item #219431 "I'm eating," which has the following poor evaluation:

```
first 5 singular values { 2.54832, 2.26775, 2.00000, 2.00000, 1.76350 }
role accuracy:             9 / 11  = 0.8182   COG-ST/ARG1
const-type precision:      2 / 4   = 0.5000
const-type recall:         2 / 2   = 1.0000
const-value accuracy:      1 / 2   = 0.5000   _eat_v_1/pron_rel
var-subtype accuracy:      2 / 8   = 0.2500
variable precision:        1 / 4   = 0.2500
variable recall:           1 / 8   = 0.1250
```

Compare this to item #219609, "He bought it and he visited his friend." where $w[0]$= 4.87846:

```
first 5 singular values: { 4.87846, 4.87484, 4.79583, 3.34596, 3.31827 }
role accuracy:            54 / 60  = 0.9000   SPECI/TENSE, COG-ST/ARG1, L-
HNDL/INDEX, R-HNDL/NUM, C-ARG/LTOP
const-type precision:     18 / 19  = 0.9474
const-type recall:        18 / 19  = 0.9474
const-value accuracy:      7 / 19  = 0.3684   m-or-f/m, +/past, ex-
ist_q_rel/pron_rel, _buy_v_1/pron_rel, _and_c/pron_rel,
_visit_v_1/pron_rel, _friend_n_1/pron_rel
var-subtype accuracy:     17 / 35  = 0.4857
variable precision:        8 / 24  = 0.3333
variable recall:           8 / 35  = 0.2286
```

Also shown in the run output file (and seen above annotating certain lines) are the names of items that failed to map. In general, the alignment of MRS variables evaluates poorly, but this could be an artifact of the *all-or-nothing* approach to the variable signature which is detailed in Section 5.2. This is even more strongly suggested when noting that the overall accuracy of just the variable sub-type , 0.76308, is much higher than the variable precision and recall scores of 0.42229 and 0.41286.

I computed the overall evaluation averages and these are shown below. These results wildly surpass my expectations for the technique. It is possible to see this and become excited just from the alignment maps which are shown, and which first hinted at the success of this technique. For example, in the 1en-1th output file, the relations of the source (left) side always appear in ascending order, but there is often considerable reshuffling visible in the index numbers shown for the target side. The role accuracy in particular shows that, with neither *a priori* knowledge of the input structures, nor manual supervision, the singular value decomposition is able to automatically obtain quality MRS alignments.

[5] Unlike as with evaluating variables proper, sub-type accuracy is per $\langle relation, role \rangle$ tuple (same as role names).

```
w[0] min: 2.54832 (219431)  max: 4.87846 (219609)
Averages over 39 single-alignment instances
             role accuracy:            0.94197
             const-type precision:     0.96088
             const-type recall:        0.98605
             const-value accuracy:     0.32646
             var-subtype accuracy:     0.76308
             variable precision:       0.42229
             variable recall:          0.41286
```

7 Summary

I presented a new technique for the unsupervised alignment of MRSes, including their variable properties or other arbitrary substructure. The method is deterministic and theoretically sound. Evaluation of this preliminary work exceeds expectations, with role and constant evaluations near perfect. There is much promise in this approach to MRS isomorphism.

This work was conducted in the context of research into bidirectional Thai-English analytical machine translation. As noted, a key obstacle in this effort has been the lack of a probabilistic parse selection model for Thai. Future work will examine using the technique described here to bootstrap a forest of Thai training derivations based on MRS isomorphism between the English Resource Grammar and the Thai grammar, when parsing sentence translation pairs.

References

Bender, E. M., Flickinger, D., & Oepen, S. (2002). The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In P*roceedings of the 2002 workshop on Grammar engineering and evaluation-Volume 15* (pp. 1-7). Association for Computational Linguistics.

Copestake, A., Flickinger, D., Pollard, C., & Sag, I. A. (2005). Minimal recursion semantics: An introduction. *Research on Language and Computation*, 3(2-3), 281-332.

Dridan, R., & Oepen, S. (2011). Parser evaluation using elementary dependency matching. In *Proceedings of the 12th International Conference on Parsing Technologies* (pp. 225-230). Association for Computational Linguistics.

Flickinger, D. (2000). On building a more efficient grammar by exploiting types. Natural Language Engineering, 6(1), 15-28.

Fujita, S., Bond, F., Oepen, S., & Tanaka, T. (2010). Exploiting semantic information for HPSG parse selection. *Research on Language and Computation*, 8(1), 1-22.

Gene H. Golub and Charles F. Van Loan. (1996). *Matrix computations (3rd ed.)*. Johns Hopkins University Press, Baltimore, MD, USA.

April Kontostathis and William M. Pottenger. (2002). *Transitivity and the co-occurrence relation in LSI*. Technical Report LU-CSE-02-005, Lehigh University.

Oepen, S., & Lønning, J. T. (2006). Discriminant-based MRS banking. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Hinrich Schütze. (1992). Dimensions of meaning. In *Proceedings of Supercomputing*.

Slayden, G. (2010) *Array TFS storage for unification grammars*. University of Washington Master's Thesis, 2010.