Lexical semantic typologies from bilingual corpora — A framework

Steffen Eger
Department of Computer Science / Carnegie Mellon University
5404 Gates Hillman Complex / Pittsburgh, PA 15213, USA
seger@cs.cmu.edu

Abstract

We present a framework, based on Sejane and Eger (2012), for inducing lexical semantic typologies for groups of languages. Our framework rests on lexical semantic association networks derived from encoding, via bilingual corpora, each language in a common reference language, the *tertium comparationis*, so that distances between languages can easily be determined.

1 Introduction

Typologocial classifications have a long tradition in linguistics. For example, typologies based on syntactic categories have been proposed e.g. by Greenberg (1961), leading a.o. to 'word order' categorizations of natural languages as belonging to SVO, VSO, etc. types. Relatedly, genealogical classification systems based on phonological and morphological similarities date back at least to the comparatists of the nineteenth centuries, among them Jacob Grimm (1785-1863), Rasmus Rask (1787-1832), and Karl Verner (1846-1896). Typological investigations into (lexical) semantic relations across languages have, in contrast, attracted little attention. Still, some results have been established such as classifications based upon treatment of animal concepts and corresponding meat concepts (see the excellent introduction to lexical typologies by Koch, 2001). As further exceptions, based on computational principles, may be considered Mehler et al. (2011), who analyze conceptual networks derived from the Wikipedia topic classification systems for different languages; Gaume et al. (2008), who propose (but do not realize, to the best of our knowledge) to compare distances between selected word pairs such as *meat/animal*, *child/fruit*, *door/mouth* across language-specific monolingual dictionaries in order to categorize the associated languages and, partly, Cooper (2008), who computes semantic distances between languages based on the curvature of translation histograms in bilingual dictionaries.

Recently, Sejane and Eger (2012) have outlined a novel approach to establishing semantic typologies based upon the language-specific polysemy relation of lexical units which entails language-dependent 'lexical semantic association networks'. To illustrate, French *bœuf* has two meanings, which we may gloss as 'cow' and 'beef' in English. Similarly, French *langue* and Spanish *lingua* mean both 'language' and 'tongue', whereas Chinese *huà* means both 'language' and 'picture'. Sejane and Eger's (2012) key idea is then that this language-specific polysemy can be made observable via the translation relation implied e.g. by a bilingual dictionary. For instance, using a Chinese-English dictionary, one might be able to uncover the polysemy of *huà* by assessing its two English translations, as given above. More formally, one might create a link (in a network) between two English words if they have a common translation in Chinese (cf. Eger and Sejane, 2010); doing the same with a Spanish-English and French-English dictionary, one would obtain three different lexical semantic association networks, all encoded in the English language, the *tertium comparationis* or *reference language* in this case. In the English networks based upon Spanish

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 90–94,
Montréal, Canada, June 7-8, 2012.

and French — Sejane and Eger (2012) call these networks the Spanish and French *versions* of English, respectively — 'language' and 'tongue' would have a link, whereas in the Chinese version of English, 'language' and 'picture' would have a link (see also Figure 1 where we illustrate this idea for English and Latin versions of German). Then, comparing these networks across languages may allow establishing a typology of lexical semantic associations.

In the current paper, we deliberate on Sejane and Eger's (2012) idea, suggesting ways to adequately formalize their approach (Section 2) and propose data sources suitable for their framework (Section 3). Moreover, in Section 4 we shortly discuss how network versions of a given reference language can be formally contrasted and suggest solutions for the *tertium comparationis* problem. In Section 5, we conclude.

2 Formal approach to lexical semantic association networks

We propose the following mathematical framework for representing lexical semantic association networks. Given n languages $L_1, \ldots, L_n$, $n \geq 2$, plus a selected reference language R distinct from $L_1, \ldots, L_n$, and bilingual translation operators $T_1, \ldots, T_n$, where T_i, $i = 1, \ldots, n$, maps (or, translates) from language L_i to the reference language R, create network graphs

$$G_i = (V_i, E_i)$$

with

$$V_i = W[R],$$

and

$$E_i = \{(u, v) \,|\, u, v \in V_i, uT_ix, xT_iv \text{ for some } x \in W[L_i]\},$$

where by $W[L]$ we denote the words of language L and by aT_ib we denote that a translates into b under T_i; moreover, we assume T_i to be symmetric such that the G_i's may be considered undirected graphs.

To generalize this a bit, we may consider *weighted graphs* where for network i, $i = 1, \ldots, n$, V_i is as above, $E_i = \{(u, v) \,|\, u, v \in V_i\}$, and each edge $(u, v) \in E_i$ has weight (being a function of)

$$d_i(u, v) = |\{x \,|\, uT_ix,\ xT_iv\}|. \quad (1)$$

Then, if u and v have no common translation x, $d_i(u, v) = 0$ and generally $d_i(u, v)$ counts the number of common translations x between u and v, entailing a generalization of the setting above, which may allow for a more fine-grained analysis and may be of importance for example for outlining semantic many-to-one relationships between a language L_i and the reference language R.

3 Possible data sources

Sejane and Eger (2012) conduct a preliminary study of their approach on the open-source bilingual dictionaries *dicts.info* (http://www.dicts.info/uddl.php). The disadvantage with using bilingual dictionaries is of course that they are scarcely available (and much less *freely* available); moreover, for the above described semantic association networks, it may be of crucial importance to have *comparable* data sources; e.g. using a general-purpose dictionary in one case and a technical dictionary in the other, or using dictionaries of vastly different sizes may severely affect the quality of results.[1]

We more generally propose to use bilingual corpora for the problem of inducing semantic association networks, where we particularly have e.g. sentence-aligned corpora like the Europarl corpus (Koehn, 2005) in mind (see also the study of Rama and Borin (2011) on cognates, with Europarl as the data basis). Then, translation relations T_i may be induced from these corpora by applying a statistical machine translation approach such as the Moses toolkit (Koehn et al., 2007). The translation relations may thus be probabilistic instead of binary, which may either be resolved via thresholding or by modifying Equation (1) as in

$$d_i(u, v) = \sum_{x \in W[L_i]} \frac{\Pr[uT_ix] + \Pr[xT_iv]}{2}$$

or

$$d_i(u, v) = \sum_{x \in W[L_i]} \Pr[uT_ix] \cdot \Pr[xT_iv],$$

both of which have (1) as special cases.

[1] As another aspect, Sejane and Eger (2012) concluded that the sizes and partly the qualities of their bilingual dictionaries were, throughout, not fully adequate for their intentions.

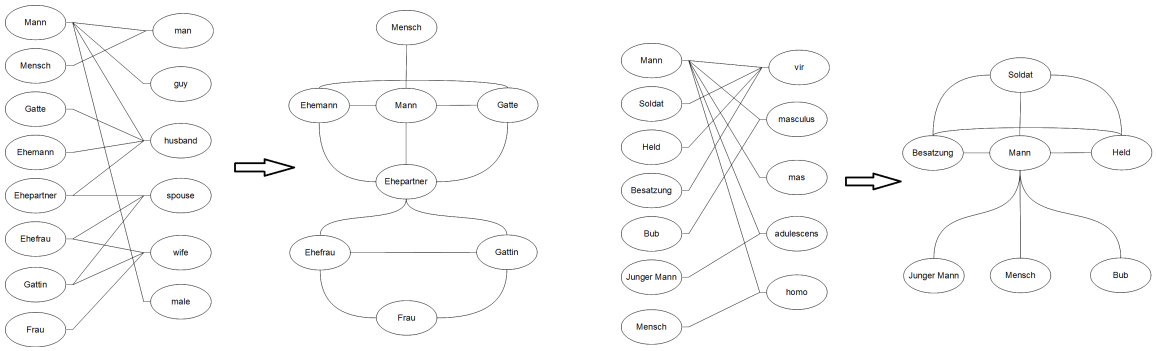

Figure 1: Bilingual dictionaries German-English and German-Latin and induced lexical semantic association networks, English and Latin versions of German. Note the similarities and differences; *Mann* 'man' and *Mensch* 'human' have a link in both versions but there is a path between *Mann* and *Frau* 'woman' only in the English version of German, whereas there exists e.g. a path between *Mann* and *Held* 'hero' only in the Latin version. Reprinted from Sejane and Eger (2012).

Using the Europarl corpus would both address the problem of size and comparability raised above; moreover, corpora may better reflect actual language use than dictionaries, which oftentimes document idiosyncractic, normative or assumed language conditions. A problem with the Europarl corpus is that it covers just a very small (and selected) subset of the world's languages, whereas it might be of particular interest for (semantic) typology to contrast large, heterogeneous classes of languages.

4 Network distance measures and the problem of *tertium comparationis*

In order to be able to induce a semantic typology from the above described lexical semantic association networks, a distance metric δ on network graphs is required,[2] that is, a function δ that maps network graphs G_i, G_j, $1 \leq i, j \leq n$, to numbers

$$\delta_{ij} = \delta(G_i, G_j) \in \mathbb{R}.$$

Such distance measures may be derived from general network statistics such as the *number of edges*, the *diameters* of the networks, network *density*, *graph entropy* via information functionals (cf. Dehmer, 2008) or *clustering coefficients* (cf. Watts and Strogatz, 1998). We believe, however, that such abstract measures can be useful only for a preliminary examination of the data. A more in-depth analysis should be based on comparing individual network vertices in two versions of the reference language. For example, we could ask about the lexical semantic difference between French and Chinese with respect to the lexical unit 'language'. One way of realizing such an analysis would be by making use of *shortest distances* between network vertices. To be more precise, let G_i and G_j be two lexical semantic network versions of a reference language R. Assume that G_i and G_j have the same number, N, of vertices, with the same labels (i.e. names of vertices such as 'language'). Let u_k, $1 \leq k \leq N$, be the k-th vertex in both graphs, with identical label across the two graphs. Moreover, let $s_i(u_k)$ and $s_j(u_k)$ be vectors whose l-th component, $1 \leq l \leq N$, is given as the shortest distance between vertex u_k and vertex u_l in graphs G_i and G_j, respectively,

$$\big(s_i(u_k)\big)_l = \text{shortest distance between } u_k \text{ and } u_l \text{ in } G_i,$$

and analogously for $s_j(u_k)$. We could then define the difference between network version G_i and G_j with respect to vertex u_k as e.g. the Euclidean distance between these two vectors,

$$\|s_i(u_k) - s_j(u_k)\|.$$

However, as useful as shortest distances may be, they do not seem to fully capture the topological structure of a network. For example, they do not indicate whether there are many or few (short) paths between two vertices, etc. (see also the discussion

[2] In this context, we identify languages with their lexical semantic association networks.

in Gaume et al., 2008). Therefore, we propose a Page-rank like (see Brin and Page, 1998; Gaume and Mathieu, 2012) procedure to compare network vertices of networks G_i and G_j. To this end, let $p_i(u_k)$, a vector of dimension N, denote the probability distribution that if, starting from vertex u_k, one may reach any of the other vertices of network G_i (and analogously for network G_j), under the following rules. In each step, starting at vertex u_k, with probability α, a 'random surfer' on the network G_i may pass from its current vertex v to any of v's neighbors with equal probability (if there are no neighbors, the surfer passes to a random vertex), and with probability $(1-\alpha)$ the surfer 'teleports' to an arbitrary vertex. The probability distribution $p_i(u_k)$, for α close to 1, may then neatly represent topological properties of network G_i, from the 'perspective' of vertex u_k. On this basis, we can, as above, determine the difference between network versions G_i and G_j with respect to vertex u_k as

$$\delta_{u_k}(G_i, G_j) = \|p_i(u_k) - p_j(u_k)\|. \qquad (2)$$

Finally, we define the (global) distance between G_i and G_j as the average over all such (local) distances,

$$\delta_{ij} = \frac{1}{N}\sum_{k=1}^{N} \delta_{u_k}(G_i, G_j). \qquad (3)$$

If, as mentioned above, we have weighted graphs, we slightly modify the random surfer's behavior. Instead of passing with uniform probability from vertex v to a neighbor vertex w of v, the surfer passes to w with probability proportional to the weight between v and w; the larger the weight the higher is the probability that the surfer ends up at w.

Then, once distance metric values δ_{ij} are given, an $n \times n$ distance matrix D may be defined whose entry (i, j) is precisely δ_{ij},

$$D_{ij} = \delta_{ij}.$$

On D, standard e.g. hierarchical clustering algorithms may be applied in order to deduce a lexical semantic typology.

Finally, we address the *tertium comparationis* problem: Given a set of languages, which one should be chosen as reference language? It might be tempting to believe that the choice of the reference language should not matter much for the resulting lexical semantic association networks, but the reference language may certainly have *some* impact. For example, if English is the reference language, the Chinese version of English might not only have a link between 'language' and 'picture' but also between 'language' and 'tongue', because of the polysemy of 'tongue' in English. If, in contrast, German was the reference language, the Chinese version of German should not have a link between *Zunge* 'tongue' and *Sprache* 'language' because *Zunge*, in German, does not mean 'language' (any more).

Thus, to avoid misspecifications based on a particular choice of reference language, we propose the following. Let $L_1, \ldots, L_n, L_{n+1}$, $n \geq 2$, be $(n+1)$ languages for which bilingual translation operators $T_{A,B}$ exist for any two languages A, B from the $(n+1)$ languages. Then let the distance between languages i and j, $1 \leq i, j \leq n+1$, be defined as

$$\Delta_{ij} = \frac{1}{n-1}\sum_{R \in L \setminus \{L_i, L_j\}} \delta(G_i^R, G_j^R),$$

where by G_i^R we denote the L_i version of R, and by L we denote the set of languages $\{L_1, \ldots, L_n, L_{n+1}\}$; in other words, we specify the distance between languages i and j as the average distance over all possible reference languages, which excludes languages i and j themselves. As above, Δ_{ij} induces a distance matrix, with which clustering can be performed.

5 Conclusion

We have presented a framework for inducing lexical semantic typologies based on the idea of Sejane and Eger (2012) to represent lexical semantic spaces of different languages in a common reference language in order to be able to contrast them. We have extended Sejane and Eger's (2012) approach by giving it a solid mathematical foundation, by suggesting more suitable data bases on which to implement their study, and by outlining adequate network distance metrics on this data. Moreover, we have addressed the *tertium comparationis* problem of the choice of the reference language. In follow-up work, we intend to bring the idea to the data, from which we expect very interesting cross-lingual lexical semantic insights.

References

S. Brin, and L. Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Seventh International World-Wide Web Conference (WWW 1998)*.

M.C. Cooper. 2008. Measuring the Semantic Distance between Languages from a Statistical Analysis of Bilingual Dictionaries. *Journal of Quantitative Linguistics*, 15 (1): 1–33.

M. Dehmer. 2008. Information processing in complex networks: Graph entropy and information functionals. *Applied Mathematics and Computation* 201: 82–94.

S. Eger, and I. Sejane. 2010. Computing semantic similarity from bilingual dictionaries. In *Proceedings of the 10th International Conference on statistical analysis of textual data (JADT 2010)*: 1217–1225.

B. Gaume, K. Duvignau, and M. Vanhove. 2008. Semantic associations and confluences in paradigmatic networks. In *From Polysemy to Semantic Change: Towards a typology of lexical semantic associations*, Amsterdam: John Benjamins: 233–267.

B. Gaume, and F. Mathieu. 2012. PageRank Induced Topology for Real-World Networks. To appear.

J. H. Greenberg. 1961. Some universals of grammar with particular reference to the order of meaningful elements. In *Universals of language*, Joseph H. Greenberg (ed.), Cambridge, MA: MIT Press: 73–113.

P. Koch. 2001. Lexical typology from a cognitive and linguistic point of view. In *Language Typology and Language Universals*, Martin Haspelmath, Ekkehard Knig, Wulf Oesterreicher, and Wolfgang Raible (eds.), Berlin: Mouton de Gruyter: 1142–1178.

P. Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit 2005*.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic, June 2007.

A. Mehler, O. Pustylnikov, and N. Diewald. 2011. Geography of Social Ontologies: Testing a Variant of the Sapir-Whorf Hypothesis in the Context of Wikipedia. *Computer Speech and Language*, 25: 716–740.

T. Rama, and L. Borin. 2011. Estimating language relationships from a parallel corpus. A study of the Europarl corpus. In *NEALT Proceedings Series (NODALIDA 2011 Conference Proceedings)*: 161–167.

I. Sejane, and S. Eger. 2012. Semantic typologies from bilingual dictionaries. To appear.

D.J. Watts, S. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature* 393 (6684): 440–442.