Do multi-sense embeddings learn more senses?
An evaluation in linear translation

Márton Makrai, Veronika Lipp[*]
Institute for Linguistics
Hungarian Academy of Sciences
{ makrai.marton, lipp.veronika}@nytud.mta.hu

Figure 1: Linear translation of word senses. The Hungarian word *finom* is ambiguous between 'fine' and 'delicious'.

Abstract

We analyze whether different sense vectors of the same word form in multi-sense word embeddings correspond to different concepts. On the more technical side of embedding-based dictionary induction, we also test whether the orthogonality constraint and related vector preprocessing techniques help in reverse nearest neighbor search. Both questions receive a negative answer.

Word sense induction (WSI) is the task of discovering senses of words without supervision (Schütze, 1998). Recent approaches include multi-sense word embeddings (MSEs), i.e. vector space models of word distribution with more vectors for ambiguous words. In MSEs, each vector is supposed to correspond to a different word sense, but in practice models frequently have different sense vectors for the same word form without an interpretable difference in meaning.

In Borbély et al. (2016b), we proposed a cross-lingual method for the evaluation of sense resolution in MSEs. The method is based on the principle that words may be ambiguous to the extent to which their postulated senses translate to different words in some other language. For the translation of words, we applied the method by Mikolov et al. (2013b) who train a translation mapping from the source language embedding to the target as a least-squares regression supervised by a seed dictionary of the few thousand most frequent words. The translation of a source word vector is the nearest neighbor of its image by the mapping in the target space. In the multi-sense setting, we have translated from MSEs. (The target embedding remained single-sense.)
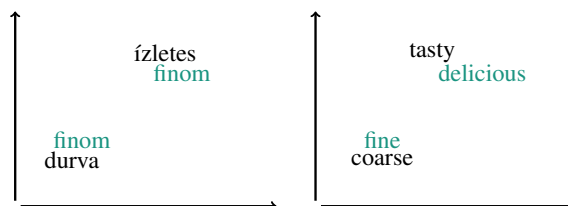
Section 1 discusses our linguistic motivation; and section 2 introduces MSEs. In section 3, we elaborate on the cross-lingual evaluation. Part of the evaluation task is to decide on empirical grounds whether different good translations of a word are synonyms or translations in different senses. Reverse nearest neighbor search, the orthogonality constraint on the translation mapping, and related techniques are also discussed. Section 4 offers experimental results with quantitative and qualitative analysis. It should be noted that our evaluation is not very strict, but rather a process of looking for something conceptually meaningful in present-day unsupervised MSE models. We make our Hungarian multi-sense embeddings[1] and the code for these experiments[2] available on the web.

1 Towards a less *delicious* inventory

We emphasize that our evaluation proposal probes an aspect of MSEs, *semantic resolution*, which is not well measured by the well-known word sense disambiguation (WSD) task that aims at classifying occurrences of a word form to different elements of a sense inventory pre-defined by some experts. Our goal in WSI is to probe the granularity of the inventory itself. The differentiation

---

[*]Veronika Lipp's contribution is section 1.1

of word senses, as already noted in Borbély et al. (2016b), is fraught with difficulties, especially when we wish to distinguish homophony, i.e. using the same written or spoken form to express different concepts, such as Russian *mir* 'world' and *mir* 'peace' from polysemy, where speakers feel that the two senses are very strongly connected, such as in Hungarian *nap* 'day' and *nap* 'sun'.

The goal of WSI can be set at two levels. We may more modestly aim to distinguish homophony from polysemy. Ideally, we could even differentiate between metonymy and metaphor, two subtypes of polysemy, discussed in more detail by Veronika Lipp in the next section.

1.1 Lexicographic Background (Veronika Lipp)

Lexical ambiguity is linguistically subdivided into two main categories: *homonymy and polysemy* (Cruse, 2004). Homonymous words have semantically unrelated and mutually incompatible meanings, such as *punch$_1$* , which means 'a blow with a fist', and *punch$_2$*, which means 'a drink'. Some have described such homonymous word meanings as essentially distinct words that accidentally have the same phonology (Murphy, 2002). Polysemous words, on the other hand, have semantically related or overlapping senses (Cruse, 2004; Jackendoff, 2002; Pustejovsky, 1995), such as *mouth* meaning both 'organ of body' and 'entrance of cave'.

Two criteria have been proposed for the distinction between homonymy and polysemy. The first criterion has to do with the *etymological* derivation of words. Words that are historically derived from distinct lexical items are taken to be homonymous. However, the etymological criterion is not always decisive. One reason is that there are many words whose historical derivation is uncertain. Another reason is that it is not always very clear how far back we should go in tracing the history of words (Lyons, 1977).

The second criterion for the distinction between homonymy and polysemy has to do with the *relatedness/unrelatedness of meaning*. The distinction between homonymy and polysemy seems to correlate with the native speaker's feeling that certain meanings are connected and that others are not. Generally, unrelatedness in meaning points to homonymy, whereas relatedness in meaning points to polysemy. However, in a large number of cases, there does not seem to be an agreement among native speakers as to whether the meanings of the words are related. So, it seems that there is not a clear dichotomy between homonymy and polysemy, but rather a continuum from "pure" homonymy to "pure" polysemy (Lyons, 1977).

Most discussions about lexical ambiguity, within theoretical and computational linguistics, concentrate on polysemy, which can be further divided into two types (Apresjan, 1974; Pustejovsky, 1995). The first type of polysemy is motivated by *metaphor (irregular polysemy)*. In metaphorical polysemy, a relation of analogy is assumed to hold between the senses of the word. The basic sense of metaphorical polysemy is literal, whereas its secondary sense is figurative. For example, the ambiguous word *eye* has the literal basic sense 'organ of the body' and the figurative secondary sense 'hole in a needle.' The other type of polysemy is motivated by *metonymy (regular polysemy)*. In metonymy, the relation that is assumed to hold between the senses of the word is that of contiguity or connectedness. In metonymic polysemy, both the basic and the secondary senses are literal. For example, the ambiguous word *chicken* has the literal basic sense referring to the animal and the literal secondary sense of the meat of that animal.

2 Multi-sense word embeddings

Vector-space language models with more vectors for each meaning of a word originate from Reisinger and Mooney (2010). Huang et al. (2012) trained the first neural-network-based MSE. Both works use a uniform number of clusters for all words that they select before training as potentially ambiguous. The first system with adaptive sense numbers and an effective open-source implementation is a modification of skip-gram Mikolov et al. (2013c), *multi-sense* skip-gram by Neelakantan et al. (2014), where new senses are introduced during training by thresholding the similarity of the present context to earlier contexts.

Bartunov et al. (2016) and Li and Jurafsky (2015) improve upon the heuristic thresholding by formulating text generation as a Dirichlet process. In `AdaGram` (Bartunov et al., 2016), senses may be merged as well as allocated during training. `mutli`-sense skip-gram[3] (Li and Jurafsky, 2015)

---

[3]Note the $l \leftrightarrow t$ metathesis in the name of the repo which is the only way of distinguishing it from the other two multi-

applies the Chinese restaurant process formalization of the Dirichlet process. `neela`, `AdaGram`, and `mutli` have a parameter for semantics resolution (more or less senses): λ, α, and γ, respectively.

MSEs are still in the research phase: Li and Jurafsky (2015) demonstrate that, when meta-parameters are carefully controlled for, MSEs introduce a slight performance boost in semantics-related tasks (semantic similarity for words and sentences, semantic relation identification, part-of-speech tagging), but similar improvements can also be achieved by simply increasing the dimension of a single-sense embedding.

3   Linear translation from MSEs

Mikolov et al. (2013b) discovered that embeddings of different languages are so similar that a linear transformation can map vectors of the source language words to the vectors of their translations.

The method uses a seed dictionary of a few thousand words to learn translation as a linear mapping $W : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ from the source (monolingual) embedding to the target: the translation $z_i \in \mathbb{R}^{d_2}$ of a source word $x_i \in \mathbb{R}^{d_1}$ is approximately its image $W x_i$ by the mapping. The translation model is trained with linear regression on the seed dictionary

$$\min_W \sum_i ||W x_i - z_i||^2$$

and can be used to collect translations for the whole vocabulary by choosing z_i to be the nearest neighbor (NN) of $W x_i$. We follow Mikolov et al. (2013b) in (i) using different metrics, Euclidean distance in training and cosine similarity in collection of translations, and in (ii) training the source model with approximately three times greater dimension than that of the target embedding.

In a multi-sense embedding scenario, Borbély et al. (2016b) take an MSE as the source model, and a single-sense embedding as target. The quality of the translation has been measured by training on the most frequent 5k word pairs and evaluating on another 1k seed pairs.

3.1   Reverse nearest neighbor search

A common problem when looking for nearest neighbors in high-dimensional spaces

_____
sense skip-gram models.

(Radovanović et al., 2010; Suzuki et al., 2013; Tomašev and Mladenic, 2013), and especially in embedding-based dictionary induction (Dinu et al., 2015; Lazaridou et al., 2015) is when there are *hubs*, data points (target words) returned as the NN (translation) of many points ($W x$s), resulting in incorrect hits (translations) in most of the cases. Dinu et al. (2015) attack the problem with a method they call *global correction*. Here, instead of the original NN, which we will call *forward* NN search to contrast with the more sophisticated method, they first rank source words by their similarity to target words. In *reverse* nearest neighbor (rNN) search, source words are translated to the target words to which they have the lowest (forward) NN rank.[4]

In reverse NN search, we restricted the vocabulary to the some tens of thousands of the most frequent words. We introduced this restriction for memory saving, because the $|V_{sr}| \times |V_{tg}|$ similarity matrix has to be sorted column-wise for forward and row-wise for reverse ranking, so at some point of the computation we keep the whole integer matrix of forward NN ranks in memory. It turned out that the restriction makes the results better: a vocabulary cutoff of $2^{15} = 32768$ both on the source and the target size yields slightly better results (74.3%) than the more ambitious $2^{16} = 65536$ (73.9%). This is not the case for forward NN search, where accuracy increases with vocabulary limit (but remains far below that of reverse NN).

3.2   Orthogonal restriction and other tricks

Xing et al. (2015) note that the original linear translation method is theoretically inconsistent due to its being based on three different similarity measures: `word2vec` itself uses the dot-product of unnormalized vectors, the translation is trained based on Euclidean distance, and neighbors are queried based on cosine similarity. They make the framework more coherent by length-normalizing the embeddings, and restricting W to preserve vector length: their matrix W is orthogonal, i.e. the mapping is a rotation. Faruqui and Dyer (2014) achieve even better results by mapping the two embeddings to a lower-dimensional bilingual space with canonical correlation analysis. Artetxe et al. (2016) analyze elements of these two works both theoretically and empirically, and

_____
[4]If more target words have the same forward rank, Dinu et al. (2015) make the decision based on cosine similarity. This tie breaking has not proven useful in our experiments.

| | | 8192 | | | | 16384 | | | | 32768 | | | |
| | | general linear | | orthogonal | | general linear | | orthogonal | | general linear | | orthogonal | |
		any	disamb	any	disamb	any	disamb	any	disamb	any	disamb	any	disamb
fwd	vanilla	28.7%	2.40%	32.1%	2.40%	36.2%	3.40%	42.0%	4.70%	36.7%	4.20%	44.5%	6.00%
	normalize	28.2%	2.20%	**33.7%**	3.40%	35.1%	2.80%	**44.4%**	5.80%	36.6%	3.80%	**48.2%**	6.00%
	+ center	26.6%	2.10%	32.8%	2.90%	32.9%	2.70%	42.0%	4.50%	34.6%	3.50%	43.9%	5.50%
rev	vanilla	**53.8%**	11.85%	51.7%	11.37%	**58.3%**	11.99%	56.6%	12.59%	**74.3%**	23.60%	73.6%	22.30%
	normalize	53.3%	11.61%	50.0%	10.90%	58.0%	12.35%	56.5%	12.59%	73.7%	24.20%	72.8%	22.10%
	+ center	51.7%	11.37%	53.3%	11.14%	57.1%	11.99%	57.7%	12.35%	69.7%	22.20%	73.5%	23.00%

Table 1: Precision@10 of forward and reverse NN translations with and without the orthogonality constraint and related techniques at vocabulary cutoffs 8192 to 32768. `any` and `disamb` are explained in section 4.3. The source has been an `AdaGram` model in 800 dimensions, $\alpha = .1$, trained on Webkorpusz with the vocabulary cut off at 8192 sense vectors.

find a combination that improves upon dictionary generation and also preserves analogies Mikolov (2013d) like

$$\mathbf{woman + king - man \approx queen}$$

among the mapped points Wx_i. They find that the orthogonality constraint is key to preserve performance in analogies, and it also improves bilingual performance. In their experiments, length normalization, when followed by centering the embeddings to $\mathbf{0}$ mean, obtains further improvements in bilingual performance without hurting monolingual performance.

4 Experiments

4.1 Data

We trained `neela`, `AdaGram` and `mutli` models on (original and stemmed[6] forms of) two semi-gigaword (.7–.8 B words) Hungarian corpora, the Hungarian Webcorpus (Webkorpusz, Halácsy et al. (2004)) and (the non-social-media part of) the Hungarian National Corpus (HNC, Oravecz et al. (2014)). We used Wiktionary as our seed dictionary, extracted with `wikt2dict`[7] (Ács et al., 2013). We tried several English embeddings as target, including the 300 dimensional skip-gram with negative sampling model `GoogleNews` released with `word2vec` (Mikolov et al., 2013a)[8], and those released with GloVe (Pennington et al., 2014)[9]. We report the best results, which were ob-

tained with the release GloVe embeddings trained on 840 B words in 300 dimensions.

4.2 Orthogonal constraint

We implemented the orthogonal restriction by computing the singular value decomposition

$$U\Sigma V = S_t^\top T_t$$

where S_t and T_t are the matrices consisting of the embedding vectors of the training word pairs in the source and the target space respectively, and taking

$$W = U\mathbf{1}V$$

where $\mathbf{1}$ is the rectangular identity matrix of appropriate shape.

Table 1 shows the effect of these factors. Precision in forward NN search follows a similar trend to that in Xing et al. (2015) and Artetxe (2016): the best combination is an orthogonal mapping between length-normalized vectors; however, centering did not help in our experiments. Reverse NNs yield much better results than the simpler method, but none of the orthogonality-related techniques give further improvement here. The cause of reverse NN's apparent insensitivity to length may be the topic of further research.

4.3 Results

We evaluate MSE models in two ways, referred to as `any` and `disamb`. The method `any` has been used for tuning the (meta)parameters of the source embedding and to choose the target: a traditional, single-sense translation has been trained between the first sense vector of each word form and its translations. (If the training word is ambiguous in the seed dictionary, all translations have been

---

[6]Follow-up work reported in section 4.5 applied a third option in preprocessing.

[7]https://github.com/juditacs/wikt2dict

[8]https://code.google.com/archive/p/word2vec/

[9]https://nlp.stanford.edu/projects/glove/

	s			covg
E	-0.04849	függő	addict, aerial	0.4
S	0.01821	alkotó	constituent, creator	0.5
S	0.05096	előzetes	preliminary, trailer	1.0
S	0.0974	kapcsolat	affair, conjunction, linkage	0.33
I	0.1361	kocsi	coach, carriage	1.0
S	0.136	futó	runner, bishop	1.0
S	0.1518	keresés	quest, scan	0.67
S	0.1574	látvány	outlook, scenery, prospect	0.6
S	0.1626	fogad	bet, greet	1.0
S	0.1873	induló	march, candidate	1.0
I	0.187	nemes	noble, peer	0.67
E	0.1934	eltérés	variance, departure	0.4
E	0.1943	alkalmazás	employ, adaptation	0.33
S	0.2016	szünet	interval, cease, recess	0.43
E	0.2032	kezdeményezés	initiation, initiative	1.0
S	0.2052	zavar	disturbance, annoy, disturb, turmoil	0.57
S	0.2054	megelőző	preceding, preventive	0.29
IE	0.2169	csomó	knot^I, lump^I, mat^E	1.0
E^5	0.21	remény	outlook, promise, expectancy	0.6
S	0.2206	bemutató	exhibition, presenter	0.67
E	0.2208	egyeztetés	reconciliation, correlation	0.5
S	0.237	előadó	auditorium, lecturer	0.67
E	0.2447	nyilatkozat	profession, declaration	0.4
I	0.2494	gazda	farmer, boss	0.67
I	0.2506	kapu	gate, portal	1.0
I	0.2515	előbbi	anterior, preceding	0.67
I	0.2558	kötelezettség	engagement, obligation	0.67
E	0.265	hangulat	morale, humour	0.5
E	0.2733	követ	succeed, haunt	0.67
SE	0.276	minta	norm^S, formula^E, specimen^S	0.75
S	0.2807	sorozat	suite, serial, succession	1.0
S	0.2935	durva	coarse, gross	0.18
I	0.3038	köt	bind, tie	0.67
E	0.3045	egyezmény	treaty, protocol	0.67
I	0.3097	megkülönböztetés	discrimination, differentiation	0.5
I	0.309	ered	stem, originate	0.5
I	0.319	hirdet	advertise, proclaim	1.0
E	0.3212	tartós	substantial, durable	1.0
I	0.3218	ajánlattevő	bidder, supplier, contractor	0.6
I	0.3299	aláírás	signing, signature	0.67
I	0.333	bír	bear, possess	1.0
I	0.3432	áldozat	sacrifice, victim, casualty	1.0
IE	0.3486	kerület	ward^I, borough^I, perimeter^E	0.3
I	0.3486	utas	fare, passenger	1.0
I	0.3564	szigorú	stern, strict	0.5
I	0.3589	bűnös	sinful, guilty	0.5
I	0.3708	rendes	orderly, ordinary	0.5
I	0.3824	eladó	salesman, vendor	0.5
I	0.3861	enyhe	tender, mild, slight	0.6
I	0.3897	maradék	residue, remainder	0.33
I	0.3986	darab	chunk, fragment	0.4
E	0.4012	hiány	poverty, shortage	0.5
I	0.4093	kutatás	exploration, quest	0.5
⋮				
I	0.4138	tanítás	tuition, lesson	0.67
I	0.4196	őszinte	frank, sincere	0.67
I	0.4229	környék	neighborhood, surroundings, vicinity	0.38
I	0.4446	ítélet	judgement, sentence	0.67
I	0.4501	gyerek	childish, kid	0.67
I	0.4521	csatorna	ditch, sewer	0.4
I	0.4547	felügyelet	surveillance, inspection, supervision	0.43
E	0.4551	ritka	rare, odd	0.5
S	0.4563	szerető	fond, lover, affectionate, mistress	0.67
I	0.4608	szeretet	affection, liking	0.67
I	0.4723	vizsgálat	inquiry, examination	0.67
I	0.4853	tömeg	mob, crowd	0.5
I	0.4903	puszta	pure, plain	0.22
I	0.4904	srác	kid, lad	1.0
I	0.4911	büntetés	penalty, sentence	0.29
I	0.4971	képviselő	delegate, representative	0.67
I	0.4975	határ	boundary, border	0.67
I	0.5001	drága	precious, dear, expensive	1.0
S	0.5093	uralkodó	prince, ruler, sovereign	0.5
I	0.5097	válás	separation, divorce	0.67
I	0.5103	ügyvéd	lawyer, advocate	0.67
I	0.5167	előnyös	advantageous, profitable, favourable	1.0
I	0.5169	merev	rigid, strict	1.0
I	0.5204	nyíltan	openly, outright	1.0
I	0.5217	noha	notwithstanding, albeit	1.0
I	0.5311	hulladék	litter, garbage, rubbish	0.43
I	0.5311	szemét	litter, garbage, rubbish	0.43
I	0.5612	kielégítő	satisfying, satisfactory	1.0
E	0.5617	vicc	joke, humour	1.0
I	0.5737	szállító	supplier, vendor	1.0
I	0.5747	óvoda	nursery, daycare, kindergarten	1.0
I	0.5754	hétköznapi	mundane, everyday, ordinary	0.75
I	0.5797	anya	mum, mummy	1.0
I	0.5824	szomszédos	neighbouring, neighbour	0.4
E	0.5931	szabadság	liberty, independence	1.0
I	0.6086	lelkész	pastor, priest	0.4
I	0.6304	fogalom	notion, conception	1.0
I	0.6474	fizetés	salary, wage	0.67
I	0.6551	táj	landscape, scenery	1.0
I	0.6583	okos	clever, smart	0.67
I	0.6707	autópálya	highway, motorway	0.5
I	0.6722	tilos	prohibited, forbidden	1.0
I	0.6811	bevezető	introduction, introductory	1.0
I	0.7025	szövetség	coalition, alliance, union	0.75
I	0.7065	fáradt	exhausted, tired, weary	1.0
I	0.7066	kiállítás	exhibit, exhibition	0.67
I	0.7135	hirdetés	advert, advertisement	1.0
I	0.7147	ésszerű	rational, logical	1.0
I	0.7664	logikai	logic, logical	1.0
I	0.7757	szervez	organise, organize, arrange	1.0
I	0.8122	furcsa	strange, odd	0.4
I	0.8277	azután	afterwards, afterward	0.67
I	0.8689	megbízható	dependable, reliable	0.67

Table 2: Hungarian words with the rNN@1 translations of their sense vectors. The first column is a post-hoc annotation by András Kornai (*E* error in translation, *I* identical, *S* separate meanings), s is the cosine similarity of the translations, and *covg* denotes the coverage of the @1 translations over all gold (good) translations.

^{5} The basic translations *hope* is missing

	dim	α/γ	p	m	any	disamb
`HNC`	800	.02		100	48.5%	7.6%
`neela` Wk	300	–	2	big	54.0%	12.4%
`HNC` stem	800	.05		big	55.1%	10.4%
`HNC`	160	.05	3	200	62.2%	15.0%
`mutli` Wk	300	.25		71	62.9%	17.4%
`Webkorpusz`	800	.05		100	65.9%	17.4%
`HNC`	600	.05	5	100	68.6%	16.6%
`HNC`	600	.1	3	50	69.1%	18.8%
`Webkorpusz`	800	.1		100	73.9%	23.9%

Table 3: Our measures, `any` and `disamb`, for different MSEs. The source embedding has been trained with `AdaGram`, except for when indicated otherwise (`neela`, `mutli`). The meta-parameters are *dim*ension, the resolution parameter (α in `AdaGram` and γ in `mutli`), the maximum number of $p$rototypes (sense vectors), and the vocabulary cutoff ($m$in-freq, the two models with *big* have practically no cut-off).

	any	disamb
`AdaGram`	73.3%	18.53%
`mutli` sense vectors	71.0%	19.46%
`mutli` context vectors	69.9%	20.76%

Table 4: The resolution trade-off between translation precision and sense distinctiveness. The source models are 600-dimensional Hungarian models trained on the de-glutinized version of the Hungarian National Corpus. Other meta-parameters have been set to default.

included in the training data.) Exploiting the multiple sense vectors, one word can have more than one translation. During the test, a source word was accepted if `any` of its sense vectors had at least one good translation among its k reverse nearest neighbors (rNN@k).

In `disamb`, we used the same translation matrix as in `any`, and inspected the translations of the different sense vectors to see whether the vectors really model different senses rather than synonyms. The lowest requirement for the non-synonymy of sense vectors s_1, s_2 is that the sets of corresponding good rNN@k translations are different. The ratio of words satisfying this requirement among all words with more than one sense vector is shown as `disamb` in table 3.

The values in Table 3 are low. This can in part be due to that the `neela` and the `mutli` models were trained with lower dimension than the best-performing model, so results here are not comparable among these different architectures. Follow-up experiments (conducted after the paper review) are reported in section 4.5.

Table 2 shows the successfully disambiguated words sorted by the cosine similarity s of good rNN@1 translations of different sense vectors. (We found that most of the few cases when there are more than two sense vectors with a good rNN@1 translation are due to the fact that the seed dictionary contains some non-basic translation, e.g. *kapcsolat* 'relationship, conjunction' has 'affair' among its seed translations. In these cases,

we chose two sense vectors arbitrarily.) Relying on s is similar to the monolingual setting of clustering the sense vectors for each word, but here we restrict our analysis to sense vectors that prove to be sensible in linear translation.

We see that most words with $s < .25$ are really ambiguous from a standard lexicographic point of view, but the translations with $s > .35$ tend to be synonyms instead.

4.4 Part of speech

The clearest case of homonymy is when unrelated senses belong to different parts of speech (POSs), and the translations reflect these POSs, e.g. *nő* 'woman; increase' or *vár* 'wait; castle'.[10] In purely semantic approaches, like `4lang` (Kornai, in press; Kornai et al., 2015), POS-difference alone is not enough for analyzing a word as ambiguous, e.g. we see the only difference between the noun and participle senses of *alkalmazott*, 'employee; applied' as *employ*ment being the *appli*cation of people for work; in the case of *belső* 'internal; interior', the noun refers to the part of a building described by the adjective.

More interesting are word forms with related senses in the same POS, e.g. *cikk*, 'item; article' (an article is an item in a newspaper); *eredmény*, 'score; result' (a score is a result measured by a number); *magas*, 'tall; high' (tall is used for people rather than high); or *idegen*, 'strange, alien; foreign', where the English translations are special cases of 'unfamiliar' (person versus language).

4.5 Follow-up experiments

After the compilation of the Festschrift, we trained models that enable a more fair comparison of

---

[10]We note that some POSs in Hungarian have blurred borders, e.g. it is debatable whether the nominal *önkéntes* 'voluntary; volunteer' is ambiguous for its POS.

`AdaGram` and `mutli` in terms of semantic resolution: we trained 600-dimensional models for Hungarian to have the 2:1 ratio between the source and the target dimension that has been reported to be optimal for this task (Mikolov et al., 2013b; Makrai, in preparation). This time we used the de-glutinized version (Borbély et al., 2016a; Nemeskey, 2017) of the Hungarian National corpus for better morphological generalization.

We can see in table 4 that there is a trade-off between the two measures, which may be interpreted to indicate that the more specific a vector is, the easier it is to translate, but if the vectors are too specific, then the translations may coincide. [11]

As a direction for future research, the analysis of the observed and inferred number of word senses as a function of word frequency may shed more light on how good a model of word ambiguity the Dirichlet Process is.

5 Acknowledgments

1957 was an influential year in linguistics: Harris (1957) developed the frequency-aware variant of the distributional method, Osgood et al. (1957) pioneered vector space models, and the author of a more recent conceptual meaning representation framework (Kornai, 2010, in press) was born. Fifty years later (more precisely in fall 2006) I met András during a class he taught on the book he was writing (Kornai, 2007). I heard about *deep cases* and *kārakas* sooner than I did about *thematic roles*. He has since taught me computational linguistics and mathematical linguistics in a master and disciple fashion.

Laozi says that a good leader does not leave a footprint, and András encouraged us to be independent and effective. One of his remarkable citations is that "It's easier to ask forgiveness than it is to get permission". The proverb is sometimes attributed to the Jesuits, who are similar to András in having had a great impact on what I've become in the past ten years. The real source of the proverb is Grace Hopper, a US navy admiral who invented the first compiler. This paper is a step in my learning to be so effective as the sources mentioned above.

András Kornai, besides the work already acknowledged, rated each item in table 2. I would like to thank the anonymous reviewer for detailed critique, both substantial and linguistic, Mátyás Lagos for reviewing language errors, and Gábor Recski and Bálint Sass for their useful comments. The orthogonal approximation was implemented following a code[12] by Gábor Borbély.

References

Judit Ács, Katalin Pajkossy, and András Kornai. 2013. Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*. Association for Computational Linguistics, Sofia, Bulgaria, pages 52–58.

Ju. D. Apresjan. 1974. Regular polysemy. *Linguitics* 142.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. 2016. Breaking sticks and ambiguities with adaptive skip-gram. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Gábor Borbély, András Kornai, Dávid Nemeskey, and Marcus Kracht. 2016a. Denoising composition in distributional semantics. In *DSALT: Distributional Semantics and Linguistic Theory*. Poster.

Gábor Borbély, Márton Makrai, Dávid Márk Nemeskey, and András Kornai. 2016b. Evaluating multi-sense embeddings for semantic resolution monolingually and in word translation. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. Association for Computational Linguistics, pages 83–89.

Alan D. Cruse. 2004. *Meaning in Language*. Oxford Textbooks in Linguistics. Oxford University Press, Oxford.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *ICLR 2015, Workshop Track*.

---

[11]There are two `mutli` models because Skip-gram and the related MSE models represent each word with two vectors, u and v in the formula $p(w_i \mid w_j) \propto \exp(u_i^\top v_j)$, that `mutli` calls *sense* versus *context* vectors respectively.

[12]https://github.com/hlt-bme-hu/eval-embed

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *EACL*. Association for Computational Linguistics, pages 462–471.

Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. 2004. Creating open language resources for Hungarian. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*. ELRA, pages 203–210.

Zellig Harris. 1957. Coocurence and transformation in liguistic structure. *Language* 33:283–340.

Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*. Association for Computational Linguistics, Jeju Island, Korea, pages 873–882.

R. Jackendoff. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford scholarship online. Oxford University Press.

András Kornai. 2007. *Mathematical linguistics*. Springer.

András Kornai. 2010. The algebra of lexical semantics. In Christian Ebert, Gerhard Jäger, and Jens Michaelis, editors, *Proceedings of the 11th Mathematics of Language Workshop*, Springer, LNAI 6149, pages 174–199.

András Kornai. in press. *Semantics*. Springer Verlag.

András Kornai, Judit Ács, Márton Makrai, Dávid Márk Nemeskey, Katalin Pajkossy, and Gábor Recski. 2015. Competence in lexical semantics. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)*. Association for Computational Linguistics, Denver, Colorado, pages 165–175.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *ACL*. Long, Oral.

Jiwei Li and Dan Jurafsky. 2015. Do multi-sense embeddings improve natural language understanding? In *EMNLP*.

John Lyons. 1977. *Semantics*. Cambridge University Press, London and New York.

Márton Makrai. in preparation. *Inter-lingual meaning representations*. PhD thesis, Loránd Eötvös University.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In Y. Bengio and Y. LeCun, editors, *International Conference on Learning Representations (ICLR 2013)*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. Xiv preprint arXiv:1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013c. Distributed representations of words and phrases and their compositionality. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, Curran Associates, Inc., pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013d. 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 (NAACL-HLT 2013)*. Association for Computational Linguistics, Atlanta, Georgia, pages 746–751.

Gregory Murphy. 2002. *The big book of concepts*. MIT Press.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *EMNLP*.

Dávid Márk Nemeskey. 2017. `emMorph` a hungarian language modeling baseline. In *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017)*. Szeged, pages 91–102.

Csaba Oravecz, Tamás Váradi, and Bálint Sass. 2014. The Hungarian Gigaword Corpus. In *Proceedings of LREC 2014*.

Charles E. Osgood, George Suci, and Percy Tannenbaum. 1957. *The measurement of meaning*. University of Illinois Press.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for

word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press.

M Radovanović, A Nanopoulos, and M Ivanović. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research* 11:2487–2531.

Joseph Reisinger and Raymond J Mooney. 2010. Multi-prototype vector-space models of word meaning. In *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 109–117.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational linguistics* .

I Suzuki, K Hara, M Shimbo, M Saerens, and K Fukumizu. 2013. Centering similarity measures to reduce hubs. In *EMNLP*.

N Tomašev and D Mladenic. 2013. Hub co-occurrence modeling for robust high-dimensional knn classification. In *Proceedings of the ECML conference*. pages 643–659.

Chao Xing, Chao Liu, Dong Wang, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *NAACL*. pages 1005–1010.