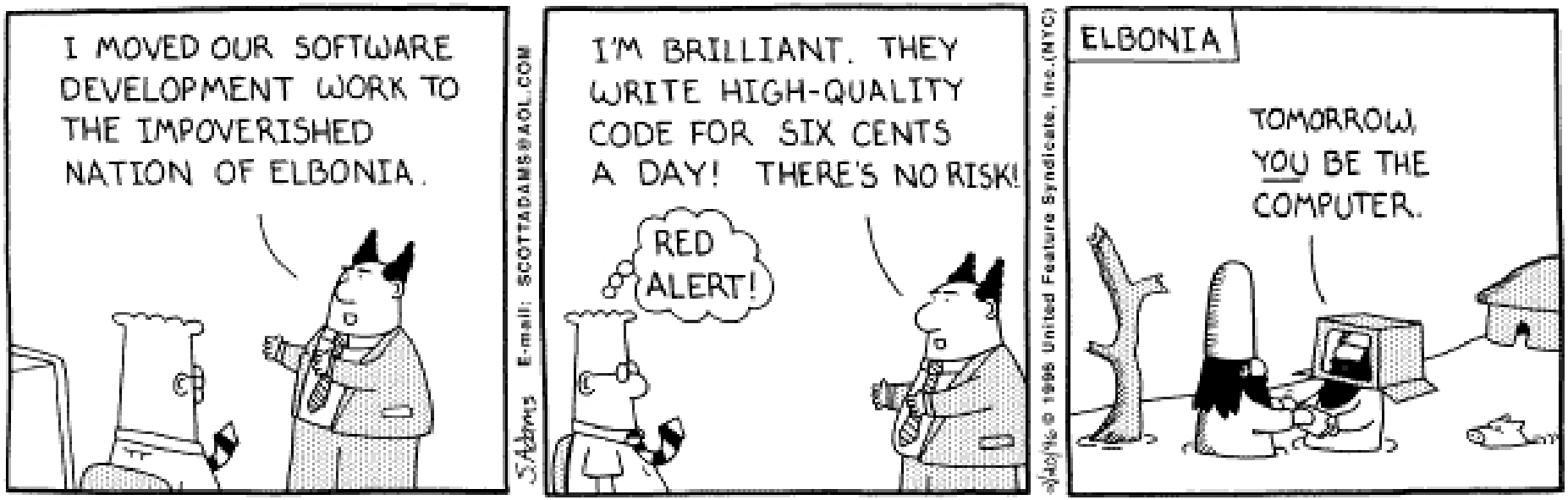
DilBERT

András Kornai
SZTAKI Computer Science Research Institute

MILAB-HLT

DIctionary Learning BERT



Plan

- Why learn dictionaries?
- How was it done before the revolution
- How could it be done with NMT
- How could it be done with language-specific BERTs
- What we have so far
- What we want

Why bother?

- Dictionaries are highly useful for humans (both beginners and pros) Proof in the pudding: people buy them, use them, create them (Wiktionary)
- Related to somewhat established shared tasks: TERMEVAL terminology extraction task Hazem 2020a, thesaurus-building Hazem 2019
- Itself a well established shared task: Bilingual Dictionary Induction Lample et al 2018 aka Bilingual Lexicon Induction Sanjanasri 2020
- Excellent abstract conceptual representations

Overview of Bengio (2013) criteria

- (B1) Continuity: if $x \sim y$ we should have $f(x) \sim f(y)$ Similarity of meaning corresponds to vector similarity
- (B2) Distributed: the data is generated by different hidden factors, and what one learns about one factor generalizes in many configurations of the other factors e.g. tense system independent of person/number system
- (B3) Hierarchical: the concepts can be defined in terms of other concepts, in a hierarchy, with more abstract concepts higher in the hierarchy, defined in terms of less abstract ones This can only be taken so far – in dictionary work we find very few irreducible primitives like *wh*

- (B4) Semi-supervised learning: With inputs X and target Y to predict, a subset of the factors explaining X's distribution explain much of Y, given X. Hence, representations that are useful for $P(X)$ tend to be useful when learning $P(Y|X)$, allowing sharing of statistical strength between the unsupervised and supervised learning tasks In a translation task this suggests learning P(target_sense|source_sense) by learning P(source_sense) first
- (B5) Shared factors across tasks: With many Ys of interest or many learning tasks in general, tasks (e.g., the corresponding $P(Y|X)$ task) are explained by factors that are shared with other tasks, allowing sharing of statistical strengths across tasks (multitask and transfer learning, domain adaptation) Words are good for tasks other than MT
- (B6) Manifolds: Probability mass concentrates near regions that have a much smaller dimensionality than the original space where the data live (autoencoders) Much more probing of the transfer vector is needed

- (B7) Natural clustering: Different values of categorical variables such as object classes are associated with separate manifolds. More precisely, the local variations on the manifold tend to preserve the value of a category, and a linear interpolation between examples of different classes in general involves going through a low-density region, i.e., $P(X|Y = i)$ for different i tend to be well separated and not overlap much. This is exploited in the manifold tangent classifier (MTC). Naming things
- (B8) Temporal and spatial coherence: Consecutive or spatially nearby observations tend to be associated with the same value of relevant categorical concepts or result in a small move on the surface of the high-density manifold. More generally, different factors change at different temporal and spatial scales, and many categorical concepts of interest change slowly This is milliscale: 20-200 msec

- (B9) Sparsity: For any given observation x, only a small fraction of the possible factors are relevant. In terms of representation, this could be represented by features that are often zero, or by the fact that most of the extracted features are insensitive to small variations of x. This can be achieved with certain forms of priors on latent variables (peaked at 0), or by using a nonlinearity whose value is often flat at 0 (i.e., 0 and with a 0 derivative), or simply by penalizing the magnitude of the Jacobian matrix (of derivatives) of the function mapping input to representation Sparsity creates structure
- (B10) Simplicity of factor dependencies: In good high-level representations, the factors are related to each other through simple, typically linear dependencies. This can be seen in many laws of physics Static word vectors come close to this ideal, especially the sparse overcomplete version

How was it done before the revolution?

1. Words explained by other words
2. Word structure divided in sequential (morphology) and parallel (senses)
3. Prague school style feature decomposition was not very successful
4. Melamed (2001) Exploiting parallel texts Implemented in HunDict
5. Classic problems: many-to-one *à cette heure* $\sim$ *now*, one-to-many, null elements
6. Segmentation still not 100% under control improved segmentation by modern methods

How could it be done with NMT?

- Leverage phrase tables already used in SMT (Gomes and Pereira Lopes 2016)
- Create cross-lingual word embeddings (Mikolov 2013, Ruder et al 2017)
- Barone (2006), Conneau et al (2018) do not even require supervision data!

  > *...we hypothesize that, if languages are used to convey thematically similar information in similar contexts, these random processes should be approximately isomorphic between languages, and that this isomorphism can be learned from the statistics of the realizations of these processes, the monolingual corpora, in principle without any form of explicit alignment (Barone 2016)*

Why DilBERT?

1. Success of unsupervised methods is more limited than appears initially (Søgaard et al 2018)
2. People are increasingly trying to align sentence and paragraph vectors before getting the word vector alignment right (Schwenk and Douze 2017)
3. The lack of fully aligned multiply parallel corpora is very limiting
4. Word-internal structure matters, but setting up universal tokenization seems hard, if not impossible
5. The greatest impediment (already at the word level) is multiple senses
6. These two considerations amount to saying that the current architecture is not set up right (ignores both sequential and parallel decomposition of words)

DilBERT objectives

1. Suffixes should have their own vector (this can be computed from kings - king $\sim$ queens - queen) B2, B3, B5, B7, B10
2. Different word senses get different vectors (multi-sense embedding, MSE) B1, B3?, B6, B7
3. First, create static (sparse) MSEs based on classic static training methods (word2vec, GloVe) B1, B10
4. (D1) Investigate relationship of monolingual BERT output (sense-disambiguated vectors) e.g. bark_1 and bark_2 to static single-sense: the assumption is that the latter is the log frequency weighted sum of the former B1, B4, B7, B9
5. (D2) develop criteria for splitting BERT cluster based on known multisense cases B3?
6. (D3) Dsambig based on neighboring context vector B4
7. Rotate monolingual DilBERTs together (technical step)
8. Evaluate on gold dictionaries as downstream task (the sales point)

Difficulties

1. Getting morpheme-like tokenization is lots of work, but should already have measurable effect on all kinds of downstream tasks (Ács et al 2021 has eval infrastructure)
2. Dealing with multi-word expressions (MWEs) is very nontrivial
3. MWE detection a good task no matter what
4. Classic embeddings (both ordinary and sparse) are easy to build on medium-size corpora (e.g. Wikipedias) but they are better on gigaword
5. Nemeskey (2020) corpus-building pipeline reasonably well oiled, but David always wants to improve it
6. Building national BERTs is more resource intensive, building them with the correct morphology will be a pain

Hopes

1. Sparse representations may be amenable to more direct inspection (XAI)
2. Lighter, faster, more understandable (big selling point)
3. Prediction: classical semantic fields (e.g. family relations, body parts, . . .) may show visible structure (selling point only for those who care)
4. Should improve embarrassing holes in SOTA systems: e.g. Google Translate has H *ima* as E. *prayer* but L. **orationis*, H *kényelmetlen* as E. *uncomfortable* but J. **fukai* 'deep, profound'
5. Should work well on bilingual dictionary induction task (core selling point)
6. Universal embedding useful in low-resource settings (another selling point)
7. May give more of a handle on POS than Lévai (2019) -style methods