CREATING SENSE VECTORS USING CROSS-LINGUAL DATA

Rachita Chhaparia, Deepanshu Gupta

October 15, 2015

Indian Institute of Technology, Kanpur

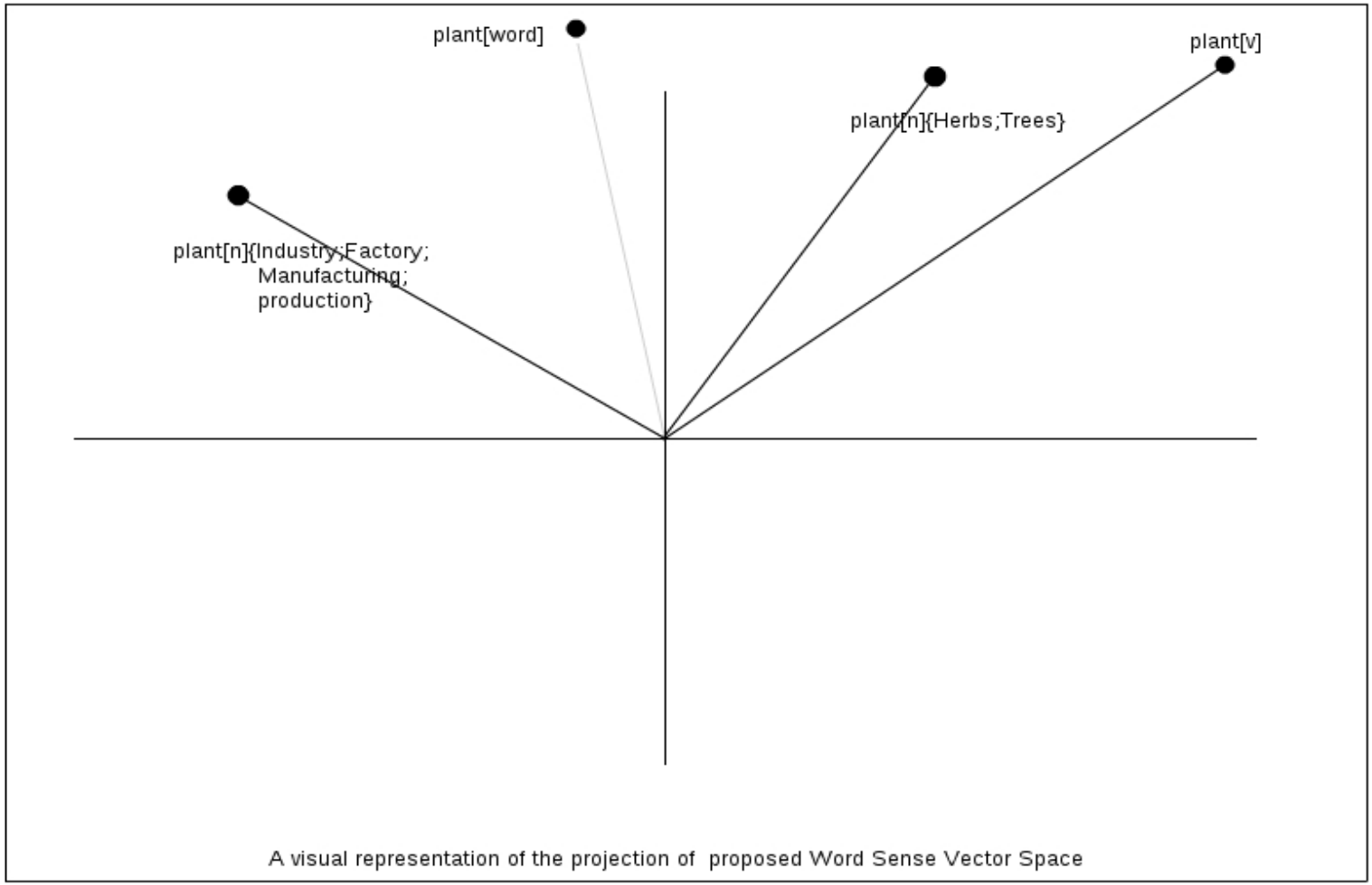
WORD SENSE DISAMBIGUATION[WSD]

Identifying and Distinguishing meaning in a context.

- Tulsi plant is well respected and worshipped in India.[Sense 2]
- TATA has set NANO plant in Gujarat.[Sense 1]
- Ram loves to plant new seeds in his garden.[Sense 3]

SENSE VECTORS

A separate vector for each sense of a polysemous word



HOW WSD AND SENSE VECTORS RELATE

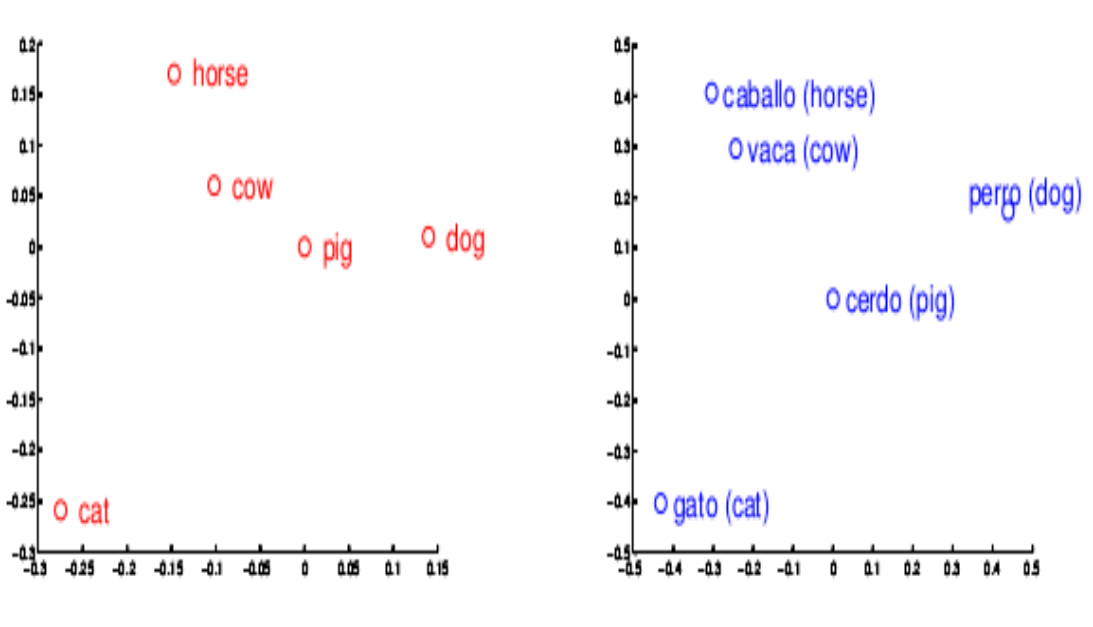
The sense vector with the highest cosine similarity with the context vectors is selected as the relevant word sense.

Example:

- The first steel plant in the southern states was established at Birmingham in 1897
- Expected similarity: $\text{plant}_1 >> \text{plant}_2 \cong \text{plant}_3$

CROSS-LINGUAL?

Inspired by ”Exploiting Similarities among Languages for Machine Translation” by Mikolov et. al.

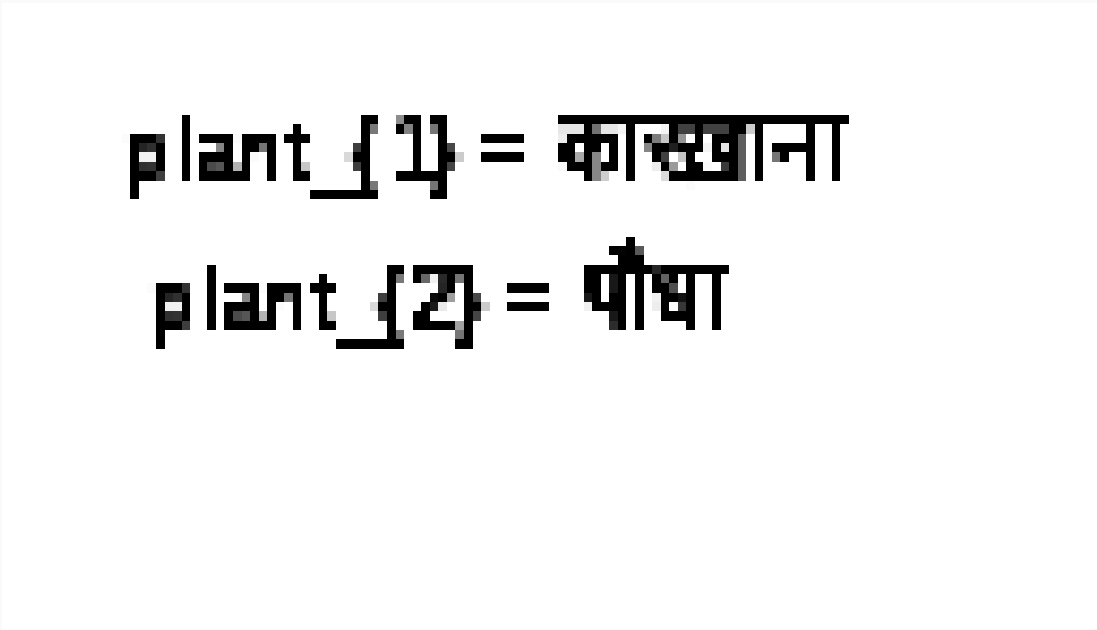


- Linear transformation between the two vector spaces
- Easy to learn: $\min_W \sum_{i=0}^{n} \|Wx_i - z_i\|$

Figure and formula taken from ”Exploiting Similarities among Languages for Machine Translation” by Mikolov et. al.

SENSE VECTORS WITH CROSS-LINGUAL DATA

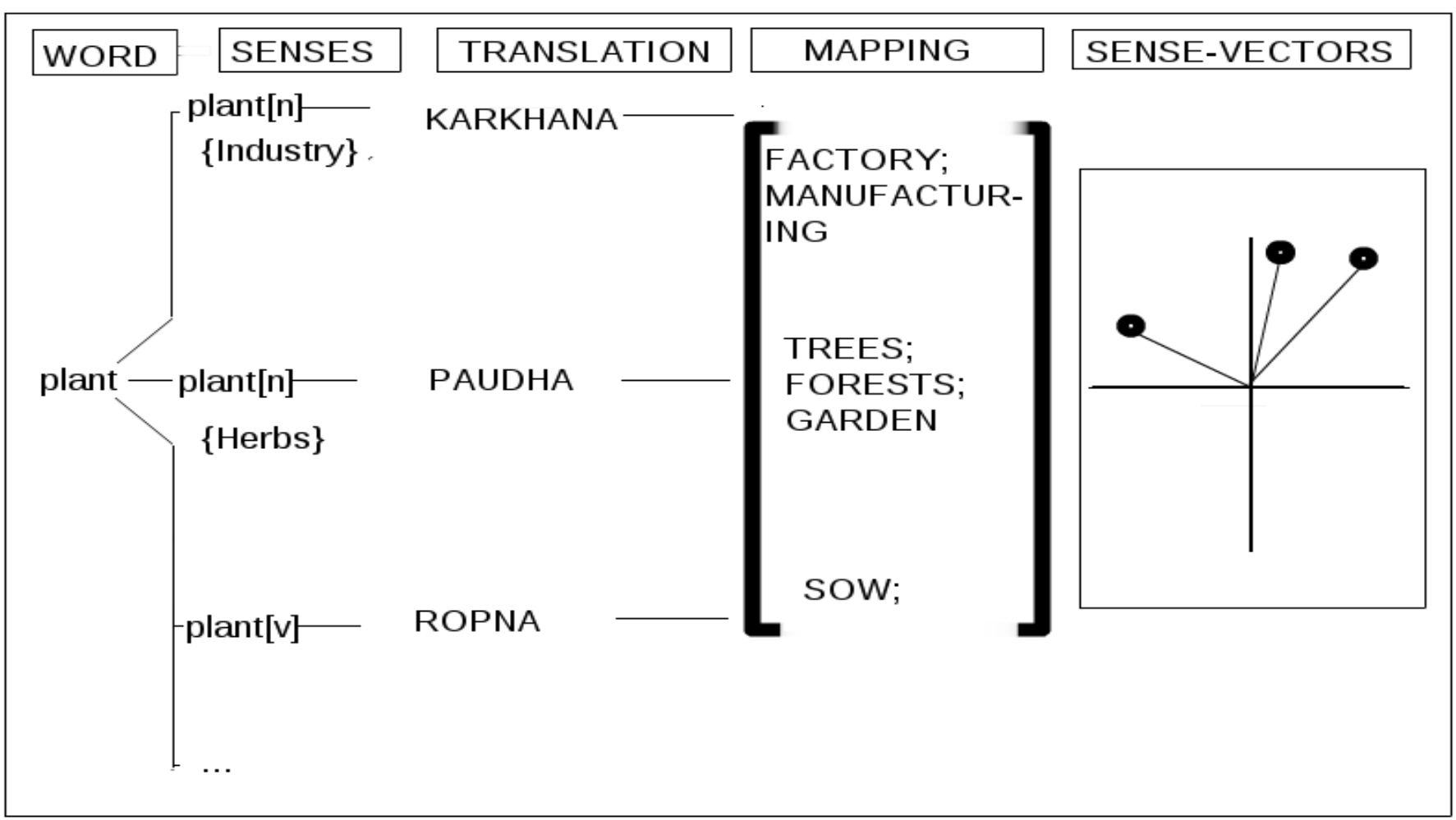
Polysemous in one language ⇒ Different monosemous words in another language



OUR METHOD

1. Create word vector space of Hindi and English: Wikipedia Corpus, word2vec
2. Learn the mapping: using bilingual dictionary of only monosemous words (collected using Google Translate)
3. Create sense vectors and disambiguate (Using Hindi WordNet IITB + Mapping learnt above)

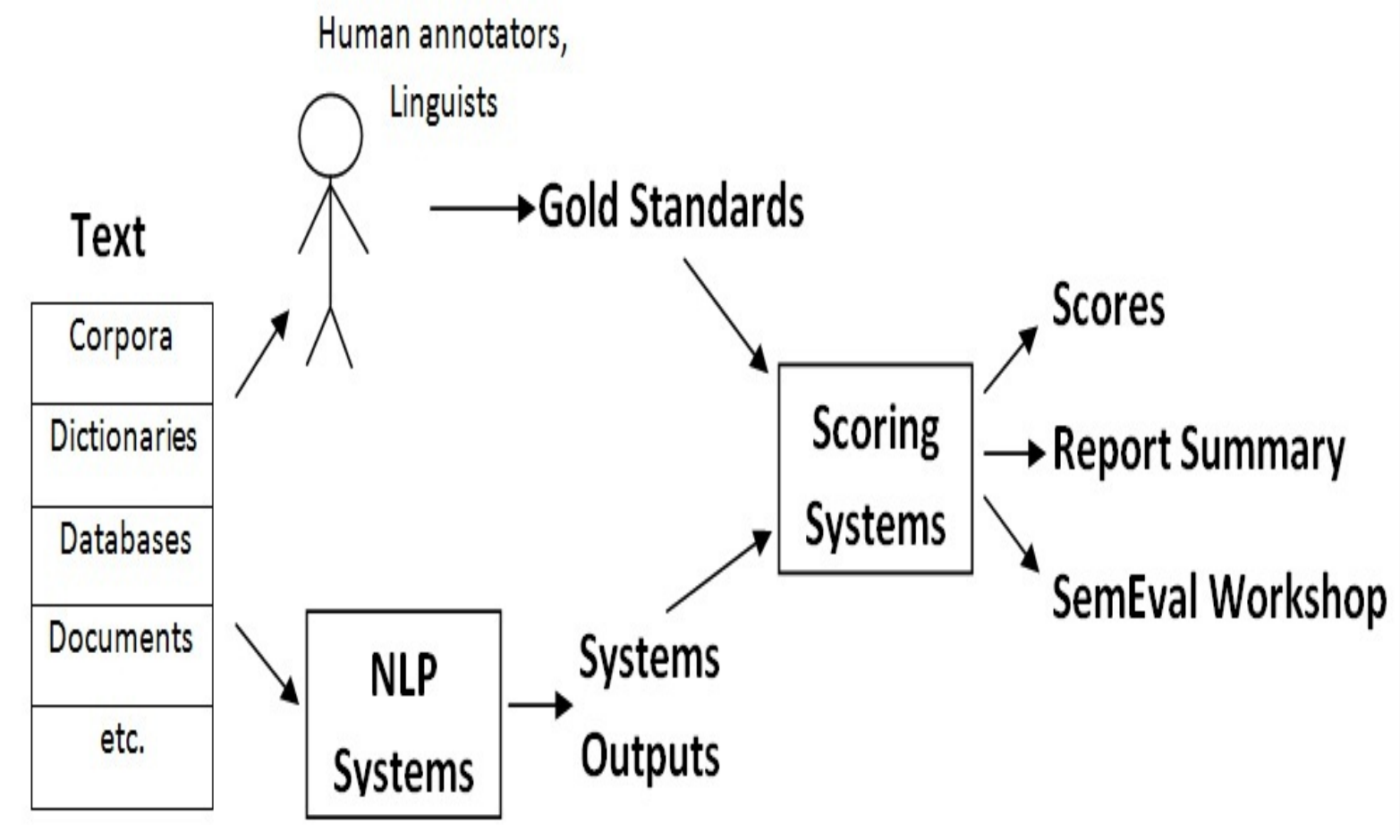
EXAMPLE



(Figure: Model Explained via plant example)

EVALUATION

Word Sense Disambiguation and Induction using Senseval 3.0 database



”SemEval framework” by alvations - Adapted from MUC introduction. Licensed under CC BY-SA 3.0

WORK

1. Hindi-English Bilingual Dictionary
2. Word embeddings for Wikipedia corpus
3. Mono-Sense word Collection
4. Mapping Code and Training
5. Word sense embedding generation
6. Evaluation

CHALLENGES EXPECTED

- Monosemous words: less frequent
- Verbs often translate to phrases in Hindi:

Plant (v) = स्थापित करना

- Polysemous in English ⇒ PolysemousinHindi :

Common (adj) = आम

Questions?