Using Unsupervised Paradigm Acquisition for Prefixes

Daniel Zeman

ÚFAL MFF, Univerzita Karlova, Praha

Morphological Paradigm

- Declension / conjugation table ⇒ set of affixes
  - German (“to have”): *ha+be, ha+st, ha+t, ha+ben, ha+bt, ha+ben, ha+tte, ha+ttest, …, hä+tte, hä+ttest, …, ge+ha+bt, …*
- Derivational morphology
  - German (“to sleep”): *schlaf+e, schläf+st, …, schlaf+end* (“sleeping”), *schlaf+end+e, schlaf+end+es, …*

Core Idea

- Assumption: 2 morphemes: stem+suffix
  - Suffix can be empty
- All splits of all words
  - (into a stem and a suffix)
- Set of suffixes seen with the same stem is a paradigm
  - In a wider sense, paradigm = set of suffixes + set of stems seen with the suffixes

Filtering 1

- Remove the paradigm if there are more suffixes than stems
  - One letter as the only stem
  - Thousands of “suffixes” – all words beginning with that letter
  - Example (en):
    - Suffixes: *…, yrup, yrups, ysop, ystem, ystem’s, …*
    - Stems: *s*

Filtering 2

- All suffixes begin with same letter ⇒ there must be another paradigm with the letter in the stems
  - Example (fi):
    - Suffixes: *a, in, ksi, lla, lle, n, na, ssa, sta* ← keep
    - Stems: *erikokoisi, funktionaalisi, logistisi, mustavalkoisi, …*
    - Suffixes: *ia, iin, iksi, illa, ille, in, ina, issa, ista*
    - Stems: *erikokois, funktionaalis, logistis, mustavalkois, …*
    - Suffixes: *sia, siin, siksi, silla, sille, sin, sina, sista*
    - Stems: *erikokoi, funktionaali, logisti, mustavalkoi, …*
    - Suffixes: *isia, isiin, isiksi, isilla, isille, isin, isina, isissa, isista*
    - Stems: *erikoko, funktionaal, logist, mustavalko, …*

Filtering 3

- If suffixes B $\subset$ A and $\forall$ C $\neq$ A : B $\not\subset$ C (if there is only one superset A of B) merge B with A (keep A)
  - Example (en):
    - Suffixes: *e, ed, er, ers, es, ing*
    - Stems: *aveng, co-manag, invad, keynot, …*
    - Superset: *e, ed, er, ers, es, es', ing*
    - Stems: *catalogu, landscap, straddl*

Superset Finding Algorithm

- Dynamic programming
- For a set of N suffixes, find all subsets sized N – 1 by dropping 1 suffix at a time
  - Mark subsets that are real paradigms as well
- Remember superset-subset links (DAG)
- Traverse the DAG sub-to-super
- If a superset is found stop at this level (find other same-sized supersets but no larger ones)
  - 69,000 English paradigms before this phase
  - 600,000 steps together constructing and querying the superset graph

Filtering 4

- Remove paradigms containing a single suffix only
- Not interesting. Group of words with the same ending. The ending may not even be a (linguistic) suffix
  - Example (en):
    - Suffix: *n*
    - Stems: *flight-inspectio, pyrennea, camerame, kufstei, …* (and thousands of others)

Paradigm Examples (en)

- Suffixes: *e, ed, es, ing, ion, ions, or*
- Stems: *calibrat, decimat, equivocat, …*

- Suffixes: *e, ed, es, ing, ion, or, ors*
- Stems: *aerat, authenticat, disseminat, …*

- Suffixes: 0, *d, r, r's, rs, s*
- Stems: *analyze, chain-smoke, collide, …*

Paradigm Examples (fi)

- Suffixes: 0, *a, an, ksi, lla, lle, n, na, ssa, sta, t*
- Stems: *asennettava, avattava, hinattava, …*

- Suffixes: *en, ksi, lla, lle, lta, n, na, ssa, sta, sti, t*
- Stems: *aatteellise, ainaise, aluepoliittise, …*

- Suffixes: *a, en, in, ksi, lla, lle, lta, na, ssa, sta*
- Stems: *ammatinharjoittaji, avustavi, jakavi, …*

Paradigm Examples (de)

- Suffixes: 0, *m, n, r, re, rem, ren, rer, res, s*
- Stems: *aggressive, bescheidene, …*

- Suffixes: 0, *e, em, en, er, es, keit, ste, sten*
- Stems: *entsetzlich, gutwillig, reichhaltig, …*

- Suffixes: 0, *m, n, r, re, ren, res, rweise, s*
- Stems: *anständige, glückliche, …*

Paradigm Examples (tr)

- Suffixes: 0, *de, den, e, i, in, iz, ize, izi, izin*
- Stems: *anketin, becerilerin, birikimlerin, …*

- Suffixes: 0, *dir, n, nde, ndeki, nden, ne, ni, nin, yle*
- Stems: *geçişleri, sürmesi, yetiştiriciliği, …*

- Suffixes: 0, *a, da, daki, dan, ı, ın, ız, ızı*
- Stems: *bakışın, baskıların, detayların, fırının, …*

Paradigm Examples (ar)

- Suffixes: 0, هم, ها, ه, نا, ك, ا
- Stems: ديون, إنفاق, أوراق, أهداف, أمور, أموال,أطفال

- Suffixes: 0, هما, هم, ها, ه, نا, ا
- Stems: نفوذ, مخاوف, قبول, جنود, تأييد, ارتياح,أسعار

- Suffixes: 0, تهما, تهم, تها, ته, تنا, ت
- Stems: عمليا, طائرا, صادرا, خلافا, تصريحا,استثمارا ...

Paradigm Examples (cs)

- Suffixes: *ou, á, é, ého, ém, ému, ý, ých, ým, ými*
- Stems: *gruzínsk, italsk, lékařsk, městsk, …*

- Suffixes: 0, *a, em, ovi, y, ů, ům*
- Stems: *divák, dlužník, obchodník, odborník, …*

- Suffixes: *a, ami, ou, u, y, ách, ám*
- Stems: *buňk, dívk, otázk, podmínk, schránk, …*

Learning Phase Outcomes

- List of paradigms
- List of known stems
- List of known suffixes
- List of stem-suffix pairs seen together

- How can we use that to segment a word?

Morphemic Segmentation

- Consider all possible splits of the word
  1. Stem & suffix known and allowed together
  2. Stem & suffix known but not together
  3. Stem is known
  4. Suffix is known
  5. Both unknown
- If there is a split where 1 or 2 holds, use it
- Otherwise, return all splits where 3 or 4 holds

Learning prefixes

- So far, just atomic stem or stem+suffix
- Now, prefix+stem+suffix (only stem must be non-empty)
- We still do not expect multiple stems (like in compounds: *jugend + welt + meister + schaft*)

Reversed Word Method

- Same algorithm but words are processed right-to-left
- Algorithm proposes “stem” and “suffix”
- Reverse them again, get prefix and $stem_2$

- This is labeled **“Zeman 3”** in the official results

Strict Prefix Segmentation

- If prefix + stem are known, remember *applicable prefix* (can be empty)
- If stem + suffix are known, remember *applicable suffix* (can be empty)
- All combinations of applicable prefixes and suffixes (and non-empty stems)
- If none are found, return dummy segmentation (just the stem)

- This is labeled **“Zeman 3”** in the official results

Rule Based Method

- Prefix = 1 to *K* first characters
- Stem = at least *L* characters
- Prefix occurs with at least *N* stems
- Stem occurs with at least *M* prefixes

- $K = 5$, $L = 2$, $M = 5$, $N = 100$

Weak Prefix Segmentation

- Take the stem-suffix segmentation found earlier
- Look for known prefix (ignore stems learned with prefixes)
- If prefix is found, make it a separate morpheme

The Hyphen Rule

- Any hyphens are replaced by morpheme boundaries
- Helps especially in English:
  - *re-creat+e, cross-examin+e, co-manag+e, free+lanc+e, -general, -in-chief, over-react, eight-page, …*

English Results

56.26	*P*	*R*	*F*
Stem+suffix	52.98	42.07	**46.90**
Rev Strict	76.92	8.47	15.27
Rule Weak	27.72	62.47	38.40

German Results

54.06	*P*	*R*	*F*
Stem+suffix	53.12	28.37	36.98
Rev Strict	72.27	7.15	13.01
Rule Weak	41.75	41.97	**41.86**

Finnish Results

48.47	*P*	*R*	*F*
Stem+suffix	58.51	20.47	30.33
Rev Strict	72.41	3.42	6.54
Rule Weak	50.12	35.85	**41.80**

Turkish Results

51.99	*P*	*R*	*F*
Stem+suffix	65.81	18.79	29.23
Rev Strict	73.30	3.01	5.79
Rule Weak	52.54	33.43	**40.86**

Arabic Results

40.87	*P*	*R*	*F*
Stem+suffix	77.24	12.73	**21.86**
Rev Strict	89.62	5.18	9.79
Rule Weak	68.96	11.20	19.27

Errors

- Noise (typos) damage results, should be recognized by word frequency
  - Example (en):
    - Suffixes: *0, ly, ness, y*
    - Stems: *abrupt, explicit*

    - Suffixes: *0, ly, ness*
    - Stems: *absent-minded, aimless, anxious, artless, assertive, …*

Errors

- Method “Rev(ersed Word) Strict” (“Zeman 3” in official results) leads to high precision and negligible recall
  - Strict segmentation is probably the responsible component here
  - Prefix examples (de):
    - Prefixes: *südo, nordo, o, südwe, nordwe, we*
    - Stems: *stprovinz, sthorizont, stchinesischen, stpolnischen, stafrikanischen, stdeutsche, …*

Rule Based Prefixes

- Very short, too frequent to be filtered
  - en: *a, a-, aa, abf, abg, ac, ag, ah, ai, ak, …*
- Real prefixes
  - en: *anti, anti-, auto, by, co, co-, dis, ex, mis, re, un, …*
  - de: *ab, an, anti, anti-, anzu, auf, aufge, aufzu, aus, ausge, be, dar, …*
- First parts of compounds
  - en: *ash, back, bank, bell, down, five-, half, …*
  - de: *abend, acht, aids, aids-, akten, alarm, alpen, …*

Future Work

- Word frequencies, filter noise (typos)
- Compounds: allow multiple stems
- Strict vs. weak segmentation: be stricter for shorter prefixes
- The naming of morphemes matters!
  - *d* and *ed* should be identical if they correspond to gold standard morpheme "PAST"

Thank + you

Dank + e

Kiito + ksi + a

Te + şekkür + ler

ش + كر + ا

Děk + uji

Tak