Unsupervised Learning of Morphology by Using Syntactic Categories

Burcu Can Suresh Manandhar

Department of Computer Science
University of York

Morpho Challenge, 2009

Outline

Morphology and Part-of-Speech (PoS)

Inspiration for another approach for morphology learning

- Correlation between morphological and syntactic information

Example

PoS category 1 : Present participles
Words : going, walking, washing . . .
PoS category 2 : Adverbs
Words : badly, deeply, strongly . . .
PoS category 3 : Plural nouns
Words : students, pupils, girls, families . . .

- Chance of joint learning of two knowledges (morphology and PoS)

Previous Research Using Morphology-PoS Together

- Hu et al. [4] extends the Minimum Description Length (MDL) based framework due to Goldsmith [3] exploring the link between morphological signatures and PoS tags
- Clark and Tim [2] experiment with the fixed endings of the words for PoS clustering
- Our work: A clustering algorithm based on PoS categories for inducing morphological paradigms

Previous Research Using Morphology-PoS Together

- Hu et al. [4] extends the Minimum Description Length (MDL) based framework due to Goldsmith [3] exploring the link between morphological signatures and PoS tags
- Clark and Tim [2] experiment with the fixed endings of the words for PoS clustering
- Our work: A clustering algorithm based on PoS categories for inducing morphological paradigms

Previous Research Using Morphology-PoS Together

- Hu et al. [4] extends the Minimum Description Length (MDL) based framework due to Goldsmith [3] exploring the link between morphological signatures and PoS tags
- Clark and Tim [2] experiment with the fixed endings of the words for PoS clustering
- Our work: A clustering algorithm based on PoS categories for inducing morphological paradigms

Outline

Inducing Syntactic Categories

Clark's [1] syntactic clustering method

Clark's [1] distributional clustering approach for syntactic categories is used.

- Each word is clustered by using its context (previous-following word)
- For the distributional similarity between the words, Kullback-Leibler (KL) divergence:

Theorem

$$D(p\|q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)} \qquad (1)$$

where p, q are the context distributions of the words being compared and x ranges over contexts.

Inducing Syntactic Categories

Clark's [1] syntactic clustering method

Clark's [1] distributional clustering approach for syntactic categories is used.

- Each word is clustered by using its context (previous-following word)
- For the distributional similarity between the words, Kullback-Leibler (KL) divergence:

Theorem

$$D(p\|q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)} \tag{1}$$

where p, q are the context distributions of the words being compared and x ranges over contexts.

Inducing Syntactic Categories

Clark's [1] syntactic clustering method

Clark's [1] distributional clustering approach for syntactic categories is used.

- Each word is clustered by using its context (previous-following word)
- For the distributional similarity between the words, Kullback-Leibler (KL) divergence:

Theorem

$$D(p\|q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)} \tag{1}$$

where p, q are the context distributions of the words being compared and x ranges over contexts.

Inducing Syntactic Categories

Clark's [1] syntactic clustering method

- In Clark's approach [1], the probability of a context for a target word is defined as:

Theorem

$$p(< w_1, w_2 >) = p(< c(w_1), c(w_2) >)p(w_1|c(w_1))p(w_2|c(w_2)) \quad (2)$$

where $c(w_1), c(w_2)$ denote the PoS cluster of words w_1, w_2 respectively.

- Starts with K clusters with most frequent words, and gradually filling with the words having the minimum KL divergence with one of the K clusters.
- We set K=77, the number of tags defined in CLAWS tagset.

Inducing Syntactic Categories

Clark's [1] syntactic clustering method

- In Clark's approach [1], the probability of a context for a target word is defined as:

Theorem

$$p(< w_1, w_2 >) = p(< c(w_1), c(w_2) >)p(w_1|c(w_1))p(w_2|c(w_2)) \quad (2)$$

where $c(w_1), c(w_2)$ denote the PoS cluster of words w_1, w_2 respectively.

- Starts with K clusters with most frequent words, and gradually filling with the words having the minimum KL divergence with one of the K clusters.
- We set K=77, the number of tags defined in CLAWS tagset.

Inducing Syntactic Categories

Clark's [1] syntactic clustering method

- In Clark's approach [1], the probability of a context for a target word is defined as:

Theorem

$$p(< w_1, w_2 >) = p(< c(w_1), c(w_2) >)p(w_1|c(w_1))p(w_2|c(w_2)) \tag{2}$$

where $c(w_1), c(w_2)$ denote the PoS cluster of words w_1, w_2 respectively.

- Starts with K clusters with most frequent words, and gradually filling with the words having the minimum KL divergence with one of the K clusters.
- We set K=77, the number of tags defined in CLAWS tagset.

Inducing Syntactic Categories

Some example PoS clusters

Some example PoS clusters are given:

Example

Cluster 1: much far badly deeply strongly thoroughly busy rapidly slightly heavily neatly widely closely easily profoundly readily eagerly . . .
Cluster 2: made found held kept bought heard played left passed finished lost changed . . .
Cluster 3: should may could would will might did does . . .
Cluster 4: working travelling flying fighting running moving playing turning . . .
Cluster 5: people men women children girls horses students pupils staff families . . .

Outline

Inducing Morphological Paradigms

Paradigm Definition

- Morphemes are tied to PoS clusters.
- Our definition of paradigm deviates from that of Goldsmith [3] in that:
  - A paradigm ϕ is a list of morpheme/cluster pairs i.e. $\phi = \{m_1/c_1, \ldots, m_n/c_n\}$.
  - Associated with each paradigm is a list of stems i.e. the list of stems that can combine with each of the morphemes m_i to produce a word belonging to the c_i PoS category.

Inducing Morphological Paradigms

Algorithm for Capturing Paradigms across PoS Clusters

Algorithm

```
Apply unsupervised PoS clustering to the input corpus
Split all the words in each PoS cluster at all split points, and create potential morphemes
For each PoS cluster c and morpheme m, compute maximum likelihood estimates of p(m | c)
Keep all m (in c) with p(m | c) > t, where t is a threshold
for all PoS clusters c1 , c2 do
    Pick morphemes m1 in c1 and m2 in c2 with the highest number of common stems
    Store φ = {m1 / c1 , m2 / c2} as the new paradigm
    Remove all words in c1 with morpheme m1 and associate these words with φ.
    Remove all words in c2 with morpheme m2 and associate these words with φ.
end for
```

Inducing Morphological Paradigms

Some Example Potential Morphemes

Table: Some high ranked potential morphemes in PoS clusters

English		German		Turkish	
Cluster	Morphemes	Cluster	Morphemes	Cluster	Morphemes
1	-s	1	-n,-en	1	-i,-si,-ri
2	-d,-ed	2	-e,-te	2	-mak,-mek,-mesi,-masi
3	-ng,-ing	3	-g,-ng,-ung	3	-an,-en
4	-y,-ly	4	-r,-er	4	-r,ar,er,-ler,-lar
5	-s,-rs,-ers	5	-n,-en,-rn,-ern	5	-r,-ir,-dir,-Ir,-dIr
6	-ing,-ng,g	6	-ch,-ich,-lich	6	-e,-a

Inducing Morphological Paradigms

Sample paradigms in English

Example

English:
ed ing : reclaim aggravat hogg trimm expell administer divert register stimulat shap rehabilitat exempt stiffen spar deceiv contaminat disciplin implement stabiliz feign mistreat extricat mimick alert seal etc
s d : implicate ditche amuse overcharge equate despise torpedoe curse plie supersede preclude snare tangle eclipse relinquishe ambushe reimburse alienate conceive vetoe waive envie negotiate diagnose etc
er ing : brows wring worship cropp cater stroll zipp moneymak tun chok hustl angl windsurf swindl cricket painkill climb heckl improvis scream scaveng panhandl lawmak bark clean lifesav beekeep toast matchmak bodybuild etc
e ed : subsid liquidat redecorat exorcis amputat fertiliz reshap regulat foreclos infring eradicat reverberat chim centralis restructur crippl rehabilitat symbolis reinstat etc
ly er : dark cheap slow quiet fair light high poor rich cool quick broad deep bright calm crisp mild clever etc
0 s : benchmark instrument pretzel wheelchair scapegoat spike infomercial catastrophe beard paycheck reserve abduction

Inducing Morphological Paradigms

Sample paradigms in Turkish

Example

Turkish:
i e : zemin faaliyetin torenler secim incelemeler eyalet nem takvim makineler yontemin becerisin gorusmeler teknigin merkezin iklim goruntuler etc
i a : cevab bakimin mektuplar esnaf olayin akisin miktar kayd yasamay bulgular sular masraflarin heyecanin kalan haklarin anlamin etc
i in : sanayiin degerlerin esin denizler duman teminat erkekler kurullarin birbirin vatandaslarimiz gelismesin milletvekillerin partisin
de e : bolgesin duzeyin yonetimin dergisin sektorun birimlerin bolgelerin tumun bolumlerin tesislerin donemin kongresin evin etc
mesi en : izlen yurutul degis uretil gerceklestiril desteklen gelistiril etc
i 0 : iman cekim mahkemelerin orneklem gaflet yazman sanat trendler mahalleler eviniz hamamlar piller ogretim olimpiyat

Inducing Morphological Paradigms

Sample paradigms in German

Example

German:
r n : kurze ehemalige eidgenoessische professionelle erste bescheidene ungewoehnliche ethnische unbekannte besondere nationalsozialistische deutsche
e en : praechtig gesichert dauerhaft bescheiden vereinbart biologisch natuerlich oekumenisch kantonal unterirdisch wissenschaftlich nahegelegen chinesisch
t en : funktionier konkurrier schneid mitwirk ansteig plaedier pfeif aufklaer schluck ausgleich weitermach abhol ankomm spazier speis aussteig aufhoer
er ung : versteiger unterdrueck erneuer vermarkt beschleunig besetz geschaeftsfuehr wirtschaftsfoerder finanzverwalt verhandl
s 0 : potential instrument flohmarkt vorhang pilotprojekt idol rechner thriller ensemble bebauungsplan empfinden defekt aufschwung

Outline

Merging Paradigms

Paradigm Merging Strategy

- For capturing more general paradigms, paradigms are merged.
- The expected paradigm accuracy to decide whether to merge two paradigms is:

$$Acc(\phi_1, \phi_2) = \frac{\frac{P}{P+N_1} + \frac{P}{P+N_2}}{2} \tag{3}$$

where ϕ_1, ϕ_2 are two paradigms, P is the number of common stems, N_1 is the number of stems in ϕ_1 that are not present in ϕ_2, and N_2 is vice-versa.

Merging Paradigms

Paradigm Merging Strategy

Algorithm

1: **for all** Paradigms ϕ_1, ϕ_2 such that $Acc(\phi_1, \phi_2) > T$, where T is a threshold **do**
2: Create new merged paradigm $\phi = \phi_1 \cup \phi_2$
3: Associate all words from ϕ_1 and ϕ_2 into ϕ
4: Delete paradigms ϕ_1, ϕ_2.
5: **end for**

Merging Paradigms

Some Example Final Paradigms After Merging - English

Example

English:
es ing e ed: sketch chew nipp debut met factor profit occurr err trudg participat necessitat stomp streak siphon stroll sprint drizzl firm climax gestur whipp roll tripp stemm dangl shuffl kindl broker chalk latch rippl collaborat chok summ propp pedal paralyz parad plough cramm slack wad saddl conjur tipp gallop totall catalogu bundl barg whittl retaliat straighten tick peek jabb slimm
s ing ed 0: benchmark mothball weed snicker thread queue jack paw yacht implement import bracket whoop conflict spoof stunt bargain honor bird fingerprint excerpt handcuff veil comment
Turkish:
u a e i : yapabileceklerin kredisin hizmetleri'n sevdikleriniz yeter' transferlerin sevkin elimiz tehlikelerin sas mucizey tehditlerin bakir muhasebesin ed gayrimenkuller ecevit' defterim izlemelerin tescilin minarey tahsilin lastikler yerlestirmey
i lar li in : ruhsat semt ikilem reaksiyonlar harc tip prim gidilmis kaldirmis degistirmis bulunmayacak aktarmis bulunacak kapanacak yazilabilecek devredilmis degisecek gelmemis
German:
er 0 e en: kassiert beguenstigt eingeholt genuegt angelastet beruehrt beinhaltet zurueckgegeben beschleunigt initiiert abgestellt bewirkt mitgenommen abgebrochen beruhigt besichtigt
0 te t er : lichtenberg limburg hill trier elmshorn dreieich praunheim heusenstamm heddernheim hellersdorf schmitt muehlheim lueneburg kassel schluechtern preungesheim rodgau bieber osnabrueck rodheim muenchen london lissabon seoul wedding treptow

Outline

Morphological Segmentation

Algorithm for Segmenting the Words

Algorithm

```
for all For each given word, w, to be segmented do
    if w already exists in a paradigm φ then
        Split w using φ as w = u + m
    else
        u = w
    end if
    If possible split u recursively from the rightmost end by
    using the morpheme dictionary as u = s_1 + ... + s_n
    otherwise s_1 = u
    If possible split s_1 into its sub-words recursively from the
    rightmost end as s_1 = w_1 + ... + w_n
end for
```

Outline

Results

Datasets Used

- We used the datasets supplied by Morpho Challenge 2009, and CLEF (Cross Language Evaluation Forum).
- CLEF datasets:
  - English: Los Angeles Times 1994 (425 mb), Glasgow Herald 1995 (154 mb).
  - German: Frankfurter Rundschau 1994 (320 mb), Der Spiegel 1994/95 (63 mb), SDA German 1994 (144 mb), SDA German 1995 (141 mb)
- For Turkish, we used a collection of manually collected newspaper archives.

Outline

Model Parameters

Prior Model Parameter Values

- Our model is unsupervised, but it requires two prior parameters to be manually set.
  - Threshold, t, on $P(m|c)$
    We set t=0.1
  - Threshold, T, on the expected accuracy of merging two paradigms
    We set T=0.75

Outline

Evaluation & Results

Competition 1 Evaluation Scores

Table: Evaluation results for English

Language	Precision	Recall	F-measure
English	58.52%	44.82%	50.76%

Evaluation & Results

Competition 1 Evaluation Scores

Table: Evaluation results for German

Language	Precision	Recall	F-measure
German - compound	73.16%	15.27%	25.27%
German - normal	57.67%	42.67%	49.05%

Evaluation & Results

Competition 1 Evaluation Scores

Table: Evaluation results for Turkish

Language	Precision	Recall	F-measure
Turkish (validity)	73.03%	8.89%	15.86%
Turkish (no validity)	41.39%	38.13%	39.70%

Conclusion & Future Work

Conclusion:

- Meaningful to use syntactic categorial information for morphology learning.
- Requires large amount of corpus for PoS clustering.
- Requires manual setting of two thresholds.

Future Work:

- Developing the current method in a probabilistic environment to get rid of the thresholds.

References I

Alexander Clark.
Inducing syntactic categories by context distribution clustering.
In *The Fourth Conference on Natural Language Learning (CoNLL)*, pages 91–94, 2000.

Alexander Clark and Issco Tim.
Combining distributional and morphological information for part of speech induction.
In *Proceedings of the 10th Annual Meeting of the European Association for Computational Linguistics (EACL)*, pages 59–66, 2003.

References II

John Goldsmith.
Unsupervised learning of the morphology of a natural language.
Computational Linguistics, 27(2):153–198, 2001.

Yu Hu, I. Matveeva, J. Goldsmith and C. Sprague.
Using morphology and syntax together in unsupervised learning.
In *Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition*, pages 20–27, June, 2005.