Case Syncretism & Disambiguating Algorithms for Urdu-Hindi POS-Tagging

Shahid Mushtaq Bhat
shahid.bhat3@gmail.com
&
Richa Shristi
rsrishti@gmail.com
Linguistic Data Consortium for Indian Languages
Central Institute of Indian Languages
Mysore, India

Outline:

- ➢ Introduction
- ➢ Case Syncretism in Urdu-Hindi
- ➢ Disambiguating Rules for Manual POS-Tagging
- ➢ Disambiguating Algorithms for Automatic POS-Tagging
- ➢ Conclusion

Introduction:

- The present work is an effort to bring into focus the key issue of Case Syncretism in Urdu-Hindi which is one of the challenges to the annotation of corpora in Indian languages both manually and automatically in terms of cognitive load to the annotator and computational complexity, respectively.

- Case Syncretism is explored from the perspective of corpus annotation, illustrating bottlenecks in the annotation process.

Continues………

What is Case?

- Case includes a variety of semantic relationships which can hold between nouns and other portions of sentence…..(Fillmore, 1968).
- Case is a system of marking dependent nouns for the type of relationships they bear to their heads (Blake, 2001:1).
- In Indian languages, it is not the diversity in the Case system that poses the actual problem in corpus annotation but the phenomenon of Case syncretism.

Continues………

What is Case-syncretism?

- Case syncretism (a typical ambiguity problem) is the mismatch between the mapping of form and function of Case markers or the postpositions.
- A single form (Postposition/Case-marker) performs various functions in more or less similar morpho-syntactic contexts.
- So, it becomes difficult to correlate the morpho-syntactic context and the function of a given form.

Case Syncretism in Urdu-Hindi:

Consider the Syncretism associated with Postposition or Case-marker se.

- It is the instrumental as well as the ablative Case-marker in Urdu-Hindi.
- Creating mapping problem (by performing multiple functions) where by forms can't be mapped on their functions i-e. lackling one to one correspondence between form and function.

- It is used for denoting:

1) Means, instrument or agency. For example: maine chAku se seb kATA (I cut the apple with a knife), usne tAr se khabar di (He sent the news by telegram).

2) The subject of the verb in in ablitative constructions and the intermediary agent of the causative constructions. For example: rAm se kAm kiyA nahin jAtA (Ram is not able to do the work), shAlu ne rAm se kAm karwAyA (Shalu made Ram do the work).

Continues………

3) Manner For example: miku dhyAn se suntA hai (Miku listens attentively), rAm mushkil se bAhar nikla (Ram came out with difficulty).

4) Cause, reason, origin. For example: chAr log pechish se mar gaye (Four people died due to dysentery), dahi doodh se bantA hai (Curd is made from milk)

5) Objects of verbs like *tell, say, ask, request* and *demand*. For example: maine rAm se poochA (I asked Ram), maine rAm se kahA (I told Ram).

6) Association. For example: rAm mohan se milA (Ram met Mohan), merA tum se koi nAtA nahin hai (I have no relation with you).

Continues………

7) Separation or going away. For example: peR se patte girte hain (The leaves fall from the tree), wo dilli se bAhar jA rahA hai (He is going out of Delhi).

8) Starting point (place or time). For example: nadi shahar se bahut door hai (The river is very far from the town), rAm somvAr se beemAr hai (Ram has been sick since Monday).

9) Difference and comparison between two. For example: rAm shyAm se lambA hai (Ram is taller than Shyam), ye kitAb usse alag hai (This book is different from that)

Continues.........

Consider the Case-syncretism associated with postposition ko.

- It is the accusative as well as the dative Case-marker in Hind-Urdu.
- Creating mapping problem by performing multiple functions.
- It is used for denoting:

1) Specificity in both animate and inanimate objects.
For example: maine laRke ko mArA (I hit the child), maine tasveer ko dekha (I saw the picture), maine ram ko kitAb di (I gave a book to Ram).

Continues………

2) Experiencer subject. For example: mujhko bhookh lagi (I felt hungry), rAm ko pencil chahiye (Ram needs a pen), mina ko maloom hai (Mina knows), laRke ko kuch yAda nahin (The boy does not remember anything).

3) Time and space. For example: tum somwAr ko Ao (You come on Monday), wo rAt ko kAm kartA hai (He works at night), wo ek tArikh ko gayA (He went on the first), shyAm shAm ko gaya (Shyam went in the evening), wo idhar ko gaya (He went this way).

4) Aspect/mood (potentiality) of the verb. For example: rAm jAne ko hai (Ram is about to go), bArish hone ko hai (It is about to rain).

Continues………

Consider the Case-syncretism associated with the postposition mein.

➢ It is used for denoting:

1) Location. For Example: rAm ghar mein hai (Ram is at home), billi boks mein ghusi (The cat entered the box).

2) Duration. For example: ye kitAb maine chAr din mein parhi (I read this book in four days), wo ek ghante mein taiyAr hua (He got ready in one hour).

3) Comparison and difference with reference to more than two. For Example: rAm in laRkon mein acchA hai (Ram is the best among these boys), bacce bacce mein farq hai (There is difference between each boy).

4) Price. For example: ye pensil das rupaye mein Ati hai (This pencil costs ten rupees), sirf itni si mithai sau rupaye mein Ayi (Only this much of sweets cost hundred rupees).

Disambiguating rules for Rules for Manual POS-Tagging:

The syncretism of 'ko' can be solved by taking into consideration the morpho-syntactic cues present in the sentence and by observing the frequency of such cues in corpora. The rules to desyncretise the Case-marker 'ko‘ are;

Rule one:

“When the Case-marker marks the direct object (DO) of the verb, the form denotes the accusative Case”.

For example:

maine laRke ko mArA (I hit the child).
mili ne tasveer ko phAri (Mili tore the picture).

Continues………

Rule two:

“When the form 'ko' marks the indirect object (IO) of the verb, it is the dative Case”.

For example:

maine ram ko kitAb di (I gave a book to Ram).

abbA ne wAhid ko paise bheje (The father sent money to Wahid).

Rule three:

“When 'ko' marks the experiencer subject, it is the dative Case”.
For Example:

mujhko bhookh lagi (I felt hungry).

laRke ko kuch yAda nahin (The boy does not remember anything).

Continues.........

Syncretism of 'se' can be solved by taking into consideration the semantic as well as syntactic cues (rarely) present in the sentence and by observing the frequency of such cues in corpora. The rules to desyncretise the Case-marker 'se’ are;

Rule one:

“When 'se' shows association between the 'se' marked nominal and the subject, it is the instrumental Case-marker”.

This association can be of various kinds, like, instrumental/agency association, manner association, causal association and comparative association.

For example:

maine chAku se seb kATA (I cut the apple with a knife) .... *Instrumental*

shAlu ne rAm se kAm karwAyA (Shalu made Ram do the work) .... *Instrumental*

miku dhyAn se suntA hai (Miku listens attentively) .... *Manner*

chAr log pechish se mar gaye (Four people died due to dysentery)....... *Causal*

rAm mohan se milA (Ram met Mohan) .... *Interactive*

rAm shyAm se lambA hai (Ram is taller than Shyam) ....*Comparative*

Continues.........

Rule two:

“When 'se' marks the subject, it is the instrumental Case-marker”.

For example:

rAm se kAm kiyA nahin jAtA (Ram is not able to do the work)

Rule three:

“When 'se' shows disassociation in time and space, it is the ablative Case-marker”.

For example:

peR se patte girte hain	(The leaves fall from the tree)
nadi shahar se bahut door hai	(The river is very far from the town)
rAm somvAr se beemAr hai	(Ram has been sick since Monday)

Continues………

Syncretism of 'mein' can be solved by taking into consideration only semantic cues present in the sentence and by observing the frequency of such cues in corpus. The rules to desyncretise the Case-marker 'mein‘ are;

Rule one:

“When 'mein' denotes location in time and space, it is the locative Case-marker”.

For example:

rAm ghar mein hai (Ram is at home)

ye kitAb maine chAr din mein parhi (I read this book in four days)

rAm in laRkon mein acchA hai (Ram is the best among these boys)

Rule two:

“When 'mein' can be substituted with the genitive marker, it is a kind of deviation from its canonical function and can be considered as an instance of the genitive Case”.

For example:

ye pensil das rupaye mein/ki Ati hai (This pencil costs ten rupees)

Disambiguating Algorithms for Automatic POS-Tagging

- As we formulated rules for manual tagging, we can also formulate some disambiguating algorithms for the automatic tagger to desyncretise the Case-markers 'ko’, ‘se’ and ‘mein’.
- However, it is worth to mention here that these algorithms cannot be implemented at the POS level because it has to take into consideration the argument structure. Once we have corpus with annotated argument structure i-e. Parsed corpus, such algorithms can be implemented.

The algorithms can be like the following :-

1) For disambiguating (ko):

“Take the string and identify the SUBJ, IO, and DO.

Identify the token either marked with 'ko' or preceded by 'ko'.

If 'ko' marks or follows the SUBJ or the IO, tag it as the Dative Case-marker.

If 'ko' marks or follows the DO, tag it as the accusative Case-marker”.

Continues.........

One thing that we can handle at the POS level is that we can at least identify the dative Case that marks the subject.

The algorithm is:

“Take the string.

Identify the N_1 marked with ‘ko’ in the given sentence.

Tag ‘ko’ as the dative Case-marker”.

For example:

laRke ko kuch yAda nahin (The boy does not remember anything).

N_1 N_2

It is important to note here that the above rule works only in the case of basic word order of Hindi-Urdu, i.e. SOV. In the corpus sentences generally occur in their canonical word order with hardly any deviation. So, being a corpus based study, this rule can be implemented at POS-Tagging level.

Continues………

2) For disambiguating (se):

“Take the string and identify the SUBJ, IO, and DO.

Identify the token either marked with 'se' or preceded by 'se'.

If 'se' marks or follows the SUBJ, IO or DO, tag it as the Instrumental Case-marker.

If 'se' is followed by a nominal modifier, tag it as the Instrumental Case-marker.

Tag the rest of the instances of 'se' as the ablative Case-marker”

Exception: This algorithm cannot capture the instances of ‘se’ marking the causee argument, manner and cause or origin due to the non-existence of syntactic cues.

-

Continues………

Like 'ko', at the POS level we can at least identify the instrumental Case that marks the subject.

The algorithm is :–

" Take the string.

Identify the N_1 marked with 'se' in the given sentence.

Tag 'se' as the instrumental Case-marker."

For example:

rAm se kAm kiyA nahin jAtA (Ram is not able to do the work).

N_1 N_2

❑ For disambiguating the Case-marker 'mein', we were unable to formulate any sort of algorithm due to absence of captureable cues.

Conclusion:

The analysis shows that it is not possible to completely capture the phenomenon of Case syncretism for automatic POS-tagging unless we take argument structure as well as semantics into consideration. Hence, we have to solve this at the higher level, i.e. at Parsing level where annotated argument structure can be used to give feed-back to the POS-Tagger to increase its efficiency.

Questions:

? ?

THANK

YOU