14

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

Shahid Mushtaqq Bhat
Richa Srishtri

Introduction: This paper is an effort to bring into focus the key issue of Case Syncretism 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. This work is based on the annotation of Hindi-Urdu corpora of 20K+ words. Here, Case Syncretism in Urdu-Hindi is explored from the perspective of corpus annotation, illustrating bottlenecks in the annotation process.

"In the past, the research on 'Case' has amounted to an examination of variety of semantic relationships which can hold between nouns and other portions of sentence....." (Fillmore,1968). According to Blake (2001:1), "Case is a system of marking dependent nouns for the type of relationships they bear to their heads." The fact is that the explicit Case marking is actually useful in establishing the semantic roles of nominals and their relationships to the verb.

In NLP works of Urdu-Hindi and other Indian languages, the explicit Case marking is one of the challenging tasks that a human annotator faces. It is not the complexity and variety in the Case system of the language that poses the problem but the phenomenon of Case syncretism. Case syncretism is defined as the mismatch between the mapping of form and function of the Case markers or the postpositions. Many times, a single form 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. It would have been possible to solve the problem by

merely identifying the morpho-syntactic peculiarities under which a given form performs a particular function and then formulating rules to be taken into account by the human annotator (such rules could have been set as default settings in the automatic tagger), but the identification of the peculiarities of the morpho-syntactic nature is very difficult as they are not always visible. However, the human-annotator can capture the semantic and pragmatic cues present in the context but how a machine (automatic tagger) can capture those cues? until the human-annotator accurately labels such peculiarities to be induced through data driven machine learning scheme or to set as default rules in Rule-based schemes of automatic POS-Tagging.

Case Syncretism in Urdu-Hindi: In Hindi-Urdu, the Case-makers are written separately from their nominal bases except in some pronominals. So, there are both the Case-markers as well as the Postpositions in Urdu-Hindi. These postpositions and the Case-markers show severe syncretism. Consider the following instances of syncretism -

'ko' is the accusative as well as the dative Case-marker in Hind-Urdu. So, a single form 'ko' performs multiple functions, creating mapping problem, whereby forms cannot be mapped properly on their functions. It is used for denoting:

1. Specificity in both animate and inanimate objects. But which one is the accusative marker and which is the dative one that will be worked out in the later section.

For example,

i) maine laRke ko mArA (I hit the child)
ii) maine tasveer ko dekha (I saw the picture)
iii) maine ram ko kitAb di (I gave a book to Ram)

2. An experiencer subject. For example,

i) mujhko bhookh lagi (I felt hungry)
ii) rAm ko pencil chahiye (Ram needs a pen)
iii) mina ko maloom hai (Mina knows)
iv) laRke ko kuch yAda nahin (The boy does not remember anything)

3. Noun denoting time and space. For example,

i) tum somwAr ko Ao (You come on Monday)
ii) wo rAt ko kAm kartA hai (He works at night)
ii) wo ek tArikh ko gayA (He went on the first)
iv) shyAm shAm ko gaya (Shyam went in the evening)
v) wo idhar ko gaya (He went this way)

4. Aspect/mood (potentiality) of the verb. For example,

i) rAm jAne ko hai (Ram is about to go)
ii) bArish hone ko hai (It is about to rain)

Now consider the Case-marker ‘se’. ‘se’ is the instrumental as well as the ablative Case-marker in Urdu-Hindi. Like 'ko', 'se' also poses mapping problem by performing multiple functions. It is used for denoting:

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

2. The subject of the verb in in abilitative constructions and the intermediary agent of the causative constructions. For example,
    i) ayAz se kAm kiyA nahin jAtA (ayAz is not able to do the work)
    ii) shAnu ne rAm se kAm karwAyA (Shanu made Ram do the work)

3. Manner. For example,
    i) miku dhyAn se suntA hai (Miku listens attentively)
    ii) rAm mushkil se bAhar nikla (Ram came out with difficulty)
4. Cause, reason, origin. For example,
    i) chAr log pechish se mar gaye (Four people died due to dysentry)
    ii) dahi doodh se bantA hai (Curd is made from milk)

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

6. Association. For example,
    i) rAm junaid se milA (Ram met Junaid)
    ii) merA tumse koi nAtA nahin hai (I have no relation with you)

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

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

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

The third instance of Case syncretism in Urdu-Hindi is that of the Case-marker ‘mein’. ‘mein’ is the locative Case-marker in Urdu-Hindi. Like 'ko' and 'se', 'mein' also poses mapping problem by performing multiple functions. It is used for denoting:

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

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

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

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

Solutions for Manual POS-Tagging:

A. 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 over a wide range of data-sets (corpora). So, we can formulate rules to desyncretise the Case-marker 'ko'.

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 phArA (Mili tore the picture)

Rule two: When the form 'ko' marks the indirect object (IO) of the verb, it is the dative Case. For example,
   i) maine ram ko kitAb di (I gave a book to Ram)
   ii) 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)

B. 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 over a wide range of data-sets. Hence, we can formulate rules to desyncretise the Case-marker 'se'.

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,

i) maine chAku se seb kATA (I cut the apple with a knife) *Instrumental*
ii) shAlu ne rAm se kAm karwAyA (Shalu made Ram do the work) *Instrumental*
iii) miku dhyAn se suntA hai (Miku listens attentively) *Manner*
iv) chAr log pechish se mar gaye (Four people died due to dysentry) *Causal*
v) maine rAm se poochA (I asked Ram) *Interactive*
vi) rAm mohan se milA (Ram met Mohan) *Interactive*
vii) rAm shyAm se lambA hai (Ram is taller than Shyam) *Comparative*

Rule two: When 'se' marks the subject, it is the instrumental Case-marker. For example,
i) 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,
i) peR se patte girte hain (The leaves fall from the tree)
i) nadi shahar se bahut door hai (The river is very far from the town)
i) rAm somvAr se beemAr hai (Ram has been sick since Monday)

C. 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 over a wide range of data-sets. Thus, we can formulate rules to desyncretise the Case-marker 'mein'.

Rule one: When 'mein' denotes location in time-period and space, it is the locative Case-marker. For example,
i) rAm ghar mein hai (Ram is at home)
i) ye kitAb maine chAr din mein parhi (I read this book in four days)
i) 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,

i) 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 in the above section, we can formulate mini-algorithms for the automatic tagger to desyncretise the Case-marker 'ko'. But it is worthwhile to mention here that these algorithms cannot be followed at the POS level because it has to take into consideration the argument structure and other higher level structures. The algorithm can be like the following-

"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."

However, 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 rule can be as following –

"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
N_1 N_2
(The boy does not remember anything)

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 of 20K+, 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.
Now, consider the mini-algorithm which de-syncretises the Case-marker 'se', though in a very limited manner. 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.

"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."

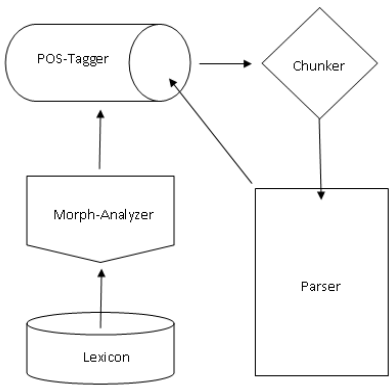
Like 'ko', at the POS level we can at least identify the instrumental Case that marks the subject. The rule can be as following –

"Identify the N_1 marked with 'se' in the given sentence
Tag 'se' as the instrumental Case-marker."

For example,
i) rAm se kAm kiyA nahin jAtA
N_1 N_2

(Ram is not able to do the work)
Regarding the Case syncretism of the locative Case-marker 'mein', it seems that it is not possible to formulate any rule for the automatic tagger. But the problem can be tackled at higher level.

Conclusion: The analysis shows that it is not possible to completely capture the phenomenon of Case syncretism for automatic tagging unless we take argument structure as well as semantics into consideration. At the POS tagging level; it is quite an impossible task to do this with whatever present techniques we have. 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 which in turn increases the efficiency of the Parser as shown in the following diagram:



Feed-back system to POS Tagger

However, what we can do at POS level is that we can keep syncretism unresolved by tagging the Post-positions/markers with ambiguous tags. For instance “ko” can be tagged as ko\Dat.Acc and “se” can be tagged as se\Ins.Abl. Finally, leaving the disambiguating task to Parsing level where argument structure is recognized and there by ambiguous tags are replaced by the discrete ones e.g. tag ko\Dat.Acc is replaced by the tag ko\Dat or ko\Acc. Similarly, other ambiguous Post-positions/markers can be disambiguated.

References:

Aronoff, Mark, and Janie Rees-Miller (eds.). 2002. *The Handbook of Linguistics*. Blackwell.
Bharati Akshar, etal. 1995. *Natural Language Processing: A Paninian Prespective.* Prentice Hall of India Private Limited:New Delhi.
Blake J., Barry. 2004. *Case*. Cambridge University Press: Cambridge.
Butt, Miriam. 2005. *Theories of Case.* CUP: Cambridge.
Butt, M. and T. H. King 2004. The Status of Case. In V. Dayal & A. Mahajan (eds.) *Clause Structure in South Asian Languages*. Springer Verlag.
Hudson, R. 1995. Does English really have case? *Journal of Linguistics* 31:375-392.
Kachru, Yamuna. 2006. *Hindi*. John Benjamins: Amsterdam/Philadelphia.
Masica, Colin. 1993. *The Indo-Aryan Languages*. Cambridge University Press: Cambridge.
Mohanan, T. 1993. Case Alternation on Objects in Hindi. *South Asian Language Review* 3:1-31.
Mohanan, T. 1994a. *Argument Structure in Hindi*. Stanford: CSLI.
Nordlinger, R. 1998. *Constructive Case.* Stanford: CSLI.
Schmidt, Ruth Laila. 1999. *Urdu: An Essential Grammar*. Routledge: London.
Sharma. Aryendra. 1994. *A Basic Grammar for Modern Hindi.* Central Hindi Directorate.
Sharma, D. 2003. Nominals and Constructive Morphology in Hindi. In M. Butt & T. H. King (eds) *Nominals Inside and Outside*. Stanford: CSLI.
Snell, R. and S. Weightman 1989. *Teach Yourself Hindi*. London: Hodder and Stoughton.
Spencer, A. 1991. *Morphological Theory.* Oxford: Blackwells.

❑❑❑