11-712: NLP Lab Report

Jonathan Barker

April 26, 2013

Abstract

This a report on the development of HindiMorph, an open source morphological analyzer for Hindi. Hindi is a morphologically rich language for which I have created an analyzer. I present a brief background on the language and the phenomena I hope to analyze. Existing work on and tools for Hindi morphology are then reviewed. After that I explain the system design and document my progress, results and ideas for future work.

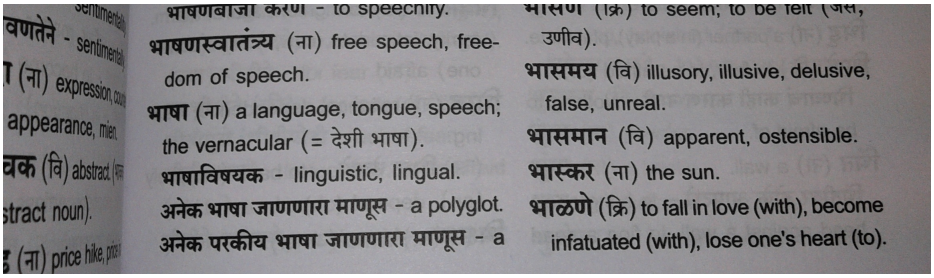
HindiMorph is an open source morphological analyzer for the Hindi language. Being a morphologically rich language, Hindi may be easier to work with when it's morphemes are tagged instead of just using the surface forms found in text. This report provides a brief introduction to the Hindi language and covers the morphological phenomina that HindiMorph hopes to analyze, in addition to summarizing previous work and other existing tools. It also documents the design of the system as well as it's development and performance.

1 Basic Information about Hindi

According to Lewis (2009), Hindi (also Khadi Boli or Khari Boli) is an Indo-European language from India that is spoken by 181,676,620 people world-wide. It is the official language of India and derives much of its formal vocabulary from Sanskrit.

Hindi is a fully developed language written in the Devanagari script. The script is written from left to right and is characterized by the long horizontal lines connecting the letters of each word. Devanagri uses spaces, making tokenization simpler. The image in Figure 1 is an example of Devanagri in a dictionary.

Figure 1: An example of Devanagri script in a dictionary.



Hindi has an SOV grammar and a rich morphology. There are two genders, two numbers, and three cases (direct, oblique, and vocative) for nouns as well as two types of nouns (type-I and type-II) within these categories. Adjectives must agree with nouns in gender, case and number athough

some do not decline at all. Verbs have 3 aspects, tense/mood and must agree in gender and number Snell (2003). There are many declensions and conjugations in Hindi, making the development of a morhpological analyzer useful.

2 Past Work on the Morphology of Hindi

3 Available Resources

[include discussion of your corpora –NAS]

4 Survey of Phenomena in Hindi

5 Initial Design

6 System Analysis on Corpus A

7 Lessons Learned and Revised Design

8 System Analysis on Corpus B

9 Final Revisions

10 Future Work

References

M. Paul Lewis, editor. *Ethnologue: Languages of the World*. SIL International, Dallas, TX, USA, sixteenth edition, 2009.

Rupert Snell. *Teach Yourself Hindi*. Teach Yourself, second edition, 2003.