Analysis and Development of Resources for Urdu Text Stemming

Abdul Jabbar

Sajid Iqbal

Muhammad Usman Ghani Khan

Department of Computer Science

Institute of Southern Punjab, Multan, Pakistan

University of Engineering & Technology, Lahore

Outline

1. **Introduction**
2. The Material used in study
3. Contribution: Resources development
4. Conclusion and Future work
5. Comments and Question Answer

Introduction (Background)

- **Morpheme:** Morpheme is the smallest grammatical unit
- **Root:** Root is a meaningful and static part of a word
- **Stem:** Stem is a morpheme or a set of concatenated morphemes that can accept an affix.
- **Affixes:** Affixes are words or letters that are attached to root words, it may be at the end or the start or both sides of the word or anywhere in the middle of the word

Related Concepts

- Urdu has both Concatenating and Non-Concatenative morphology
- Concatenating morphology
  - Concatenation is main operation to form new word
  - For Example:

خوش	خوشگواری، خوشگوار، خوشگواریت، ناخوش، ناخوشگوار، ناخوشگواری، ناخوشگواریت

- Non-Concatenative morphology
  - (main morphological operation infixation)
- For Example: معلوم سے علم, استانی سے استاد

Stemming

- Stemming is a computational procedure that reduces all the words with same root /stem
  - Connected, Connecting, Connection, and Connections ⟹ Connect
- It is the process for reducing inflected (تدابیر) and derivational words (بااخلاق) to their stem, base or root form – generally a written word form.
  - For example Urdu words تدابیر، بااخلاق in which با and ا are affixes and its stem are اخلاق and تدبیر.
- Software applications that perform stemming are known as stemmers.

Application Areas

- Used to improve retrieval effectiveness
- To reduce the size of indexing files.
- Text Mining (TM)
- Text Summarization (TS)
- Text Classification (TC)
- Sentiment Analysis (SA)
- Domain Analysis (DA)
- Computational Linguistics (CL)

Stemming Approaches

- There are three types of stemmer

1. **Rule base stemmer**
   - Knowledge of underline language is required
2. **Statistical stemmer**
   - statistical information from a large corpus of a given language to learn morphology of words
3. **Hybrid stemmer**
   - Hybrid stemmer are the combination of rule base stemmer and Statistical stemmer

Outline

1. Introduction
2. **The Material used in study**
3. Contribution: Resources development
4. Conclusion and Future work
5. Comments and Question Answer

Urdu grammar books

- **Five Urdu grammar books**
  a) "قواعد اردو" , Haq, Molvi Abdul (1996)
  b) "قواعد بنیادی اردو" , Bloch, Dr. Sohail Ahmad (2012)
  c) URDU: AN ESSENTIAL GRAMMER Schmidt, Ruth Lail (1999).
  d) "اردو قواعد وانشاء" , Board, P. T. (2010)
  e) "تخلیق اردو گرائمر" , UEP (2014)

Research Papers

- Five research papers about Urdu morphology
  a) "Morphology of the Urdu Language", (Qureshi, and Awan, 2012)
  b) "Finite-state morphological analyzer for urdu.“, Hussain (2004)
  c) "Analysis, design and implementation of Urdu morphological analyzer." Rizvi and Hussain (2005)
  d) "The morphology of loanwords in Urdu: the Persian, Arabic and English strands." Islam, (2012).

Online Resources

- Online resources included
  - BBC Urdu news
  - DAWN News
  - Two online Urdu dictionary available at
    - اردولغت (urduencyclopedia)
    - اردولغت (online Urdu dictionary)

Outline

1. Introduction
2. The Material used in study
3. **Contribution: Resources development**
4. Conclusion and Future work
5. Comments and Question Answer

Contribution 1: Affix List Development

- Prefix List (PL) 643
- Suffixes List (SL) 568

List	Suffixes
L1	و،ت،ا،ی،ے
L 2	کش،ات،یک،سا
L 3	نما،جگر،غیر
L 4	خانہ،کاری،
L 5	انگیز،ڈہال،ائیں،روائی،پزیری
L6	مزاحیہ،شاہانہ،افزائی،خداداد
L7	ترازوئے،خطیبانہ،خودداری،گزارانہ،مجرمانہ،پاشیدگی، وفاداری
L8	سپردداری،جمالیاتی

list	Prefixes
L 1	ب،ن
L 2	بن،کم،نا
L 3	ابن،غیر،ہمہ
L 4	ابدی،ابلق،اپی
L 5	ابتدائی،ابالی،ابتدائے،میثاق
L6	گلوگیر،ابنائے،معروضی
L7	آرائیاں،آہنگیوں،قرارداد،برادران،خودداری
L8	برادریوں،دستاویزی،مخالفانہ،اصطلاحات

Contribution 2: SW Collection

- Stop words are common and high frequency words(Atwan et al. 2013)
- Urdu stop words list contains postpositions, determiners, pronouns, and conjunctions (Gupta et al. 2015)
- Stop word list consisting of 1124 Urdu stop words

Urdu stop words
کا
پر
سے
تم
یہ

Contribution 3: Stem Words List (SWL)

- Urdu stem word list contains 40904

Urdu stem words
خط
حاصل
سجدہ
وارث
حبیب
وکیل
مرتبہ

Contribution 4: Rules for Infix

- 10 rules for Urdu word length 4 letters with variations of rules,
- 12 rules for Urdu word length 5 letters with variations of rules
- and 13 rules for Urdu word length 6 letters with variations of rules.

Set of Rules: Words Length 5						
Rule No. 01						
Index		**4**	**3**	**2**	**1**	**0**
Orthographic pattern		.	ا	.	.	ا
Input word	احکام	م	ا	ک	ح	ا
Stem Word	حکم	م		ک	ح	ا

Contribution 4: Rules for Infix (con't)

- Variation of Rule No. 01, words length 5

Set of Rules: Words Length 5						
Rule No. 01 Variation A						
Index		**4**	**3**	**2**	**1**	**0**
Orthographic pattern		.	ا	.	.	ا
Input word	احساس	س	ا	س	ح	ا
Invalid Stem	حسس	س		س	ح	
Deletion	حس			س	ح	
Stem Word	حس			س	ح	
Rule No. 01 Variation B						
index		**4**	**3**	**2**	**1**	**0**
Orthographic pattern		.	ا	.	.	ا
Input word	اتحاف	ف	ا	ح	ت	ا
Invalid Stem		ف		ح	ت	
insertion	تحفه		ه	ف	ح	ت
Output word	تحفه		ه	ف	ح	ت

Outline

1. Introduction
2. The Material used in study
3. Contribution: Resources development
4. **Conclusion and Future work**
5. Comments and Question Answer

Conclusion and Future work

- We developed the prefix, suffix list, stop words list, stem words list and also developed rules for infixes handling
- These resources freely available online at https://sourceforge.net/projects/resource-for-urdu-stemmer/
- We believe that these resources provide a good level of confidence to develop a robust stemmer for Urdu language.
- In future, we indented to enhance this corpus by adding other sources.

Outline

1. Introduction
2. The Material used in study
3. Contribution: Resources development
4. Conclusion and Future work
5. **Comments and Question Answer**

Any Question or Comment:

?

Thank You

References

- Haq, Molvi Abdul (1996), "قواعد اردو", Anjaman Tariqi e Urdu, New Dehli (Hind)
- Bloch, Dr. Sohail Ahmad (2012), "قواعد بنیادی اردو", Muqtadrah Qumi Zuban Pakistan, Islamabad.
- Schmidt, Ruth Lail (1999). URDU: AN ESSENTIAL GRAMMER.
- Board, P. T. (2010). "اردو قواعد وانشاء" for Class-10th. Lahore: Punjab Textbook Board.
- UEP (2014)," تخلیق اردو گرائمر", for class 8th, Unique Education Publisher, Urdu bazar Lahore.
- Qureshi, Anwar & Awan"Morphology of the Urdu Language", International Journal of Research in Linguistics and Lexicography, INTJR-Volume 1-Issue 3, September 2012,
- Hussain, Sara. "Finite-state morphological analyzer for urdu." PhD diss., National University of Computer & Emerging Sciences, 2004.
- Rizvi, SM Jafar, and Mutawarra Hussain. "Analysis, design and implementation of Urdu morphological analyzer." In Engineering Sciences and Technology, 2005. SCONEST 2005. Student Conference on, pp. 1-7. IEEE, 2005.
- Islam, Riaz Ahmed. "The morphology of loanwords in Urdu: the Persian, Arabic and English strands." (2012).
- BBC Urdu news (http://www.bbc.com/urdu)
- DAWN News (http://www.dawnnews.tv/)
- اردو لغت (urduencyclopedia), http://www.urduencyclopedia.org/urdudictionary/index.php
- اردو لغت (online Urdu dictionary), http://182.180.102.251:8081/oud/default.aspx
- Atwan, Jaffar, Masnizah Mohd, and Ghassan Kanaan. "Enhanced Arabic information retrieval: Light stemming and stop words." In Soft Computing Applications and Intelligent Systems, pp. 219-228. Springer Berlin Heidelberg, 2013.
- Gupta, V., Joshi, N., & Mathur, I. (2015, February). Design & development of rule based inflectional and derivational Urdu stemmer 'Usal'. In Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 2015 International Conference on (pp. 7-12). IEEE