

Sense Tagged CLE Urdu Digest Corpus

Saba Urooj, Sana Shams,
Sarmad Hussain, Farah Adeeba



Centre for Language Engineering
Al-Khawarizmi Institute of Computer Science
University of Engineering and Technology Lahore,
Pakistan

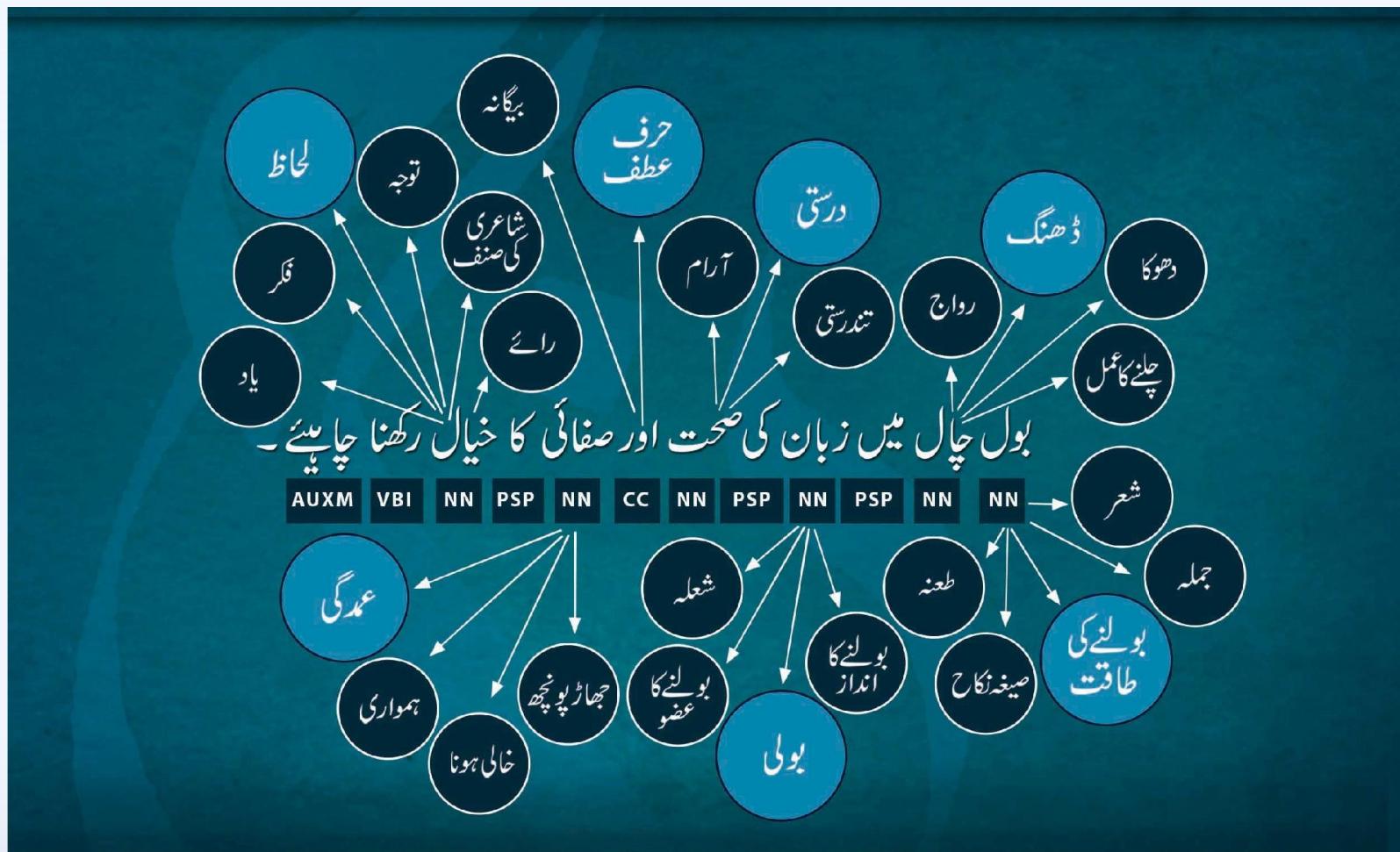
Contents

- What is WSD & Sense Tagged Corpus?
- Why It is useful?
- An overview of existing Wordnet Tagged Corpora
- The process of Developing Urdu Sense Tagged Corpus
- Current State of Urdu Sense Tagged Corpus
- Future Work

What is WSD?

- “To define the correct meaning of a word in its respective context” [1]
- A word’s meaning comes from patterns of use
- To see this usage pattern, corpora needs to be sense tagged

WSD



Why is Sense Tagged Corpus Useful?

- Natural Language Application
 - traditional and statistical machine translation
 - computer-assisted lexicography [2]
- Statistic data extracted from sense tagged corpus can be implemented in
 - Information Retrieval (IR)
 - Information Extraction
 - Text Summarization [3]

An Overview of Wordnet Tagged Corpora

- Hector (20 Million words) [4]
- English SemCor (200, 000 words) [5]
- DSO Corpus (192,800 words) [4]
- Dutch SemCor (400,000 words) [6]
- Bulgarian Word sense tagged corpus (99,480 words) [7]
- Chinese Word sense tagged corpus (60,895 words) [8]

DSO Corpus

- 192,800 words tagged
- Tagged with WordNet 1.5 senses
- It is distributed on the Linguistic Data Consortium (LDC)
- The DSO corpus uses the targeted tagging approach

Need

- For the development of a successful and accurate WSD
- Attention has been paid to develop
 - Urdu Digest Corpus [9]
 - Urdu Wordnet [10]
- Sense annotation is unexplored

Sense Tagged CLE Urdu Digest Corpus

Developing Urdu Sense Tagged Corpus

- Four linguistic resources have been used
 - Urdu Wordnet 1.0 Wordlist¹
 - CLE Urdu Digest Corpus [9]
 - Urdu WordNet [10]
 - Urdu Morphological Analyzer [11]

¹http://www.cle.org.pk/software/ling_resources/UrduWordNetWordlist.htm

Urdu Wordnet 1.0 Wordlist

5 of 108 - + Automatic Zoom

Center for Language Engineering 2013 Urdu WordNet 1.0 Wordlist

141	ارتقا	153	ارفع	165	اساطیر	177	استدلال
142	ارتقائی	154	اڑنا	166	اسامی	178	استر
143	ازٹکاز	155	ازتالیں	167	اسباب	179	استری
144	ارجمند	156	ازتینیں	168	اسپرٹ	180	استصواب
145	اردو	157	اڑنا	169	اسپیکر	181	استطاعت
146	ارز	158	ازبر	170	استناد	182	استعداد

CLE Urdu Digest Corpus

دیا کا ہر فرد کامیابی کا آرزو مند ہے۔ ناکامی سے سب بھرا تے میں۔ عزت، دولت، راحت اور عافیت کی تنگی کے سمجھی شیائی میں۔ لیکن اصل کامیابی کیا چیز ہے؟ اور حقیقی عزت و راحت کس طرح فصیب ہوتی ہے؟ اس بھیسے بہت کم لوگ واقع میں۔ اگر آپ حقیقی کامیابی کے گرد جانتا پاہتے میں تو ذاکر زادہ منیر عامر کی تازہ تصنیف آئینہ کردار پڑھیے۔^{۱۲} صفحی کی اس کتاب کا ایک ایک حرف بصیرت کے در پیچے کھولنے پر مامور ہے۔

راقم نے اس کتاب کا مطالعہ کیا تو لفظ و معنی کی لکھناں دلکش کر مسکو ہو گیا جس چیز نے فاس طور پر متاثر کیا وہ ذاکر صاحب کا فہم قرآن ہے۔ بظاہر ذاکر صاحب بخاطب یعنی وسی کے معلم ادیات میں لیکن درحقیقت وہ ایک داعی، ایک عارف، ایک محقق، ایک مرد، ایک مقرر، ایک آموزگار اخلاق اور قلم و قرطاس کے فرمزاں میں۔ فی وہی پر ان کی تقریبیں ہوئے ذوق و شوق سے سنی جاتی میں۔ ان کی باتیں عید کی بولیاں میں۔ بے انتیار دل میں اترتی ہیں جاتی میں۔ سب سے زیادہ گرناہیے غلبی یہ ہے کہ ان کی گنگوہ قرآن کریم کی بر مغل آیات اور ارشادات رسالت مآب سے پہلی بھگتی ہے میں۔

ع پر تو سے آنکاب کے ذرے میں جان بے

Urdu WordNet

urdu WordNet

English WordNet Mapping

مفهوم الفاظ

	[NOUN]
1.	[101688] دھاگوں رسیوں یا نائلون کی بنی ہوئی بڑی سی جالی جس سے مچھلیاں اور پرند وغیرہ پکڑتے ہیں { چڑیا کو مٹی کی وجہ سے جال نظر نہ آیا اور وہ اس میں پہنس گئی }
	پہندا جال دام دھوکا
2.	[101689] کسی کا بھکاوا { میں نے خبردار بھی کیا تھا لیکن تم پھر اس کے چنگل میں پہنس گئے } جال
3.	[104159] الجھی ہوئی چیز { زندگی ایک جال ہے جسے سمجھنا ناممکن ہے } جال
4.	[104184] کسی چیز کا تسلسل، سلسلہ، توسعی { پورے ملک میں نہروں کا جال پھیلایا جا رہا }

جال گیردار
جال
جام
جامع
جامہ
جان
جانا
جانب
جانبداری
جانچ
جانچنا
جاندار
جانفسانی
جاننا
جانی
رام

Urdu Morphological Analyzer

	A	B	C	D	E
1	ابا		ابالنا		اتر
2	enumerator>> string not found		enumerator>> البا		اترنا>>
3		ابل	enumerator>> الپنے		اترنے>>
4	enumerator>> البا		enumerator>> الپنی		اترنی>>
5	enumerator>> الپنے		enumerator>> الپنی		اترنی>>
6	enumerator>> الپنی		enumerator>> ابل		اتر>>
7	enumerator>> الپنی		enumerator>> البا		اترنا>>
8	enumerator>> ابل		enumerator>> الپنے		اترے>>
9	enumerator>> البا		enumerator>> الپنی		اترتی>>
10	enumerator>> الپنے		enumerator>> الپنیں		اترتبیں>>
11	enumerator>> الپنی		enumerator>> ابل		اتر>>
12	enumerator>> الپنیں		enumerator>> ابلا		اترا>>
13	enumerator>> ابل		enumerator>> الپے		اترے>>
14	enumerator>> ابلا		enumerator>> الپی		اتری>>
15	enumerator>> الپے		enumerator>> الپس		اترین>>
16	enumerator>> الپی		enumerator>> الپے		اترے>>

Sense Annotation Method

- Word forms were linked to corresponding word senses in Urdu Wordnet
- Targeted tagging approach was used
- Benefit of using this approach

Urdu Word-Sense Annotation tool (اردو مفہوم کار)

دوں صدی/103495 میں عرب سیاح، ابن حوقل نے لکھا اندلس میں خوشحالی کی اساس زراعت پر ہے، یہاں نہروں کا جال / 104184 بچہا/
 نایاب مجھلی/19 100197 گنی دو ہفتے قبل/102951 انڈونیشیا کا مجھیہ، یا سین لوامائیہ سولا ویسی (sulawisi) میں چھلیاں پکڑا 100197 رہا تھا کہ اپنکے جال / 101688 میں باسکٹ بال کی ایجاد/100693 سے تقریباً 101565 ایک/100699 100436 سال/100699 بعد لیوائیں (n) نامی ایک/100699 امریکی باشدے نے عام توکری کی بگد لوہے کا ایسا/100695 پچھلے کا جال / 104185 کھولنے کے لیے ایک/100699 ڈوری/102343 لٹکی 104322 ہوتی تھی، جب گیند/100484 جال / 104185 میں پلی جاتی تو لیفڑی ڈوری/102343 کھینچ کر گیند/104184 کے دوران موجودہ جال / 104185 100229 (t Ne) پہلی/100449 100844 استعمال/100844 ہوا۔

3

اس سے گیند/100484 جال / 104185 میں بانے کے بعد خود بخود/102070 زمین پر آ/101756 رہتی ہے۔
 موقع محل دیکھا/102287 کر جال / 101689 بھی پہنچنک/ Literary Reference & Idioms > دستی میں۔

2

SynSetID	POS	Sense	Example	Words
101688	Noun	دھاگوں رسمیں یا نکلوں کی عنی ہوئی چڑی ہی جالی جس سے چھلیاں اور پندو شیر کھوئتے ہیں	ہڈیا کو منی کی وجہ سے جال نظر آئی اور وہ اس میں پھنس گی میں نے خیردار بھی کیا تھا لیکن تم پھر اس کے بھگل میں پھنس کے	ا، جال، دام، دھوکا
101689	Noun	کسی کارکدا	ننگی ایک جان ہے جسے کھجھنا ملکن ہے	جال
104159	Noun	اچھی ہوئی پیچ	پوسے نک ک میں نہروں کا جال چھلایا جا رہا ہے	جال
104184	Noun	کسی پیچ کا سلسہ، سلسہ، تو پیچ	رسی یا ڈوری وغیرہ سے حقیقی ہوئی ہوئی جالی ہو تھت مصاد (عماکسی غسل وغیرہ) کے لیے استعمال کی جاتی ہو	جال
104185	Noun	چھلے کا جال کھولنے کے لیے ایک ڈوری لگکی ہوئی تھی، جب گیند جال میں پلی جاتی تو لیفڑی ڈوری کھینچ کر گیند کا دینا		

1

Source: Options Save

1. انج
2. بانچنا
3. جانے
4. ارسی
5. بانٹھانی

Three views of the interface

1. Selection view: displays the list of high frequency words and enables the annotator to select a target word
2. Wordnet view: displays the linguistic information available in WordNet for the selected lexical item
3. Corpus view: displays the corpus with all occurrences of the target word

Corpus annotation tags

- Exact Match
- Insufficient Context
- Literary Reference/symbolic Sense
- Non-standard Usage of language
- Word Sense Not Available
- Other

Options

104159	Noun	اچھی ہوئی چیز	
104184	Noun	کسی چیز کا سلسلہ، سلسلہ، توسعہ	
104185	Noun	اصد (تمواہ کی تکمیل وغیرہ) کے لیے استعمال کی جاتی ہے	

CLT14

The screenshot shows a software application window. At the top, there is a yellow header bar with the text "Sense Tagged CLE Urdu Digest Corpus". Below the header, the title "Corpus annotation tags" is displayed in a large, bold, black font. On the left side of the main area, there is a vertical list of annotation tags: "Exact Match", "Insufficient Context", "Literary Reference/symbolic Sense", "Non-standard Usage of language", "Word Sense Not Available", and "Other". To the right of this list, a dropdown menu is open, showing the same six options. Further down the page, there is a table with four columns. The first column contains row numbers (104159, 104184, 104185). The second column contains part-of-speech tags ("Noun"). The third column contains the annotated text in English ("اچھی ہوئی چیز", "کسی چیز کا سلسلہ، سلسلہ، توسعہ", "اصد (تمواہ کی تکمیل وغیرہ) کے لیے استعمال کی جاتی ہے"). The fourth column contains the corresponding text in Urdu. At the bottom left of the table, there is a yellow button labeled "CLT14".

Current state of Urdu sense tagged corpus

Sense Tagged Corpus	
Total no. Of sentences in the corpus	5611
Total no. of words in the corpus	100,000
Tagged total word types	2225
Tagged total sense types	2285
Tagged total word tokens	17006

	No. of senses tagged					
Words	1	2	3	4	5	6
	1522	345	118	49	21	14
	No. of senses tagged					
Words	7	8	9	10	11	12
	3	2	3	3	1	1

Improving Wordnet via sense tagging

- Consistency of sense definition with POS
 - Interpretive definition doesn't match with the POS

Words	POS	Sense	Modified
افسردہ	ADJ	تھکن یا غم اور دکھ ہونے کی وجہ سے بیزار ہو جانا	تھکن یا غم اور دکھ ہونے کی وجہ سے بیزار ہونے والا، بجھا بجھا سا

Consistency of sense example with POS

- Example associated with the synset doesn't match with POS of the word

Words	POS	Sense	Example	Modified
خزندہ	Noun	رینگنے والا جانور	خزندہ جانوروں کی بھرماں تھیں وہاں خزندوں کی بھرماں تھیں	وہاں خزندوں کی بھرماں تھیں

Consistency of definition across synset

- It is very important that all the members of the synset have equal relationship with sense meaning

POS	Sense	Example	Words	Modified
Verb	پرت والی چیز کا کوئی حصہ الگ کرنا، ٹکڑے کرنا	اس نے تختہ کے توڑ کشتنی پھاڑ دیا	پھاڑنا، چیرنا ادھیرنا	پھاڑنا، چیرنا

Addition of senses available in the corpus

- During the process of annotation all the “word sense not available” tags were reported back to the WordNet team

Challenges in the process of sense tagging

- Tool specific
 - Normalization
- Language specific
 - Non-standardized translations
 - Foreign language borrowed words
 - Complex predicates

Tool specific

- Normalization

- annotation tool was unable to match corpus in some contexts e.g. vow (و) with hamza above case (ؑ)
- these combinations were typed in different formats in corpus and WordNet and hence requiring the process of normalization
- the process of representing texts into consistent formats [23]

Language specific

- Non-standardized translations
 - It was ambiguous to tag non-standardized translations of English words which have become part of Urdu language
 - e.g. بلند فشار خون (i.e. high blood pressure and)
- Foreign language borrowed words
 - Mapping was not found for such words which are borrowed from foreign language and have been lexicalized for Urdu
 - e.g. test match, basket ball and interview

Complex Predicates

- Complex predicates
- Main Verb + Light Verb in Urdu [12]

- اس میں پائے جانے والی حیاتین ہماری آنکھوں کو
- صحتمند رکھتی ہے -

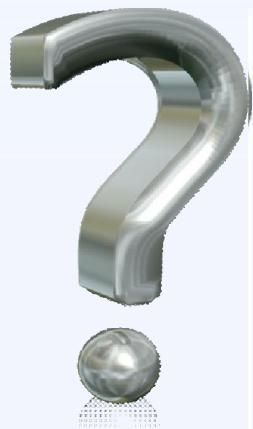
Future Work

- The current Sense tagged CLE Urdu Digest Corpus has
 - 100k Words
 - Of which 17006 are sense tagged
- This manually data can act as seed data
- Development of semantically rich corpus

Acknowledgements

- This work has been supported through the collaboration of
 - University of Konstanz, Germany
 - German Research Exchange Program (DAAD)

Thank you



References

1. A. Kilgarriff, "Word senses", *Word Sense Disambiguation*, Springer Netherlands, 2006.
2. P. Resnik, "WSD in NLP applications", *Word Sense Disambiguation: Algorithms and Applications* (2006).
3. S. J. Ker, C. R. Huang, J. F. Hong, S. Y. Liu, H. L. Jian, I. L. Su & S. K. Hsieh, "Design and Prototype of a Large-scale and Fully Sense-tagged Corpus." Large-Scale Knowledge Resources. Construction and Application. *Springer Berlin Heidelberg*, 2008
4. A. Kilgarriff, "SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Program", Proceedings of First International Conference on *Language Resources and Evaluation*, Granada, 1998.

References

5. S. Landes, C. Leacock and R. I. Tengi, “Building semantic concordances”, *WordNet: An electronic lexical database*, 1998: 199-216.
6. P. Vossen, A. Görög, R. Izquierdo, & A. van den Bosch, “DutchSemCor: Targeting the ideal sense-tagged corpus”, *LREC*, 2012.
7. S. Koeva, S. Leseva, E. Tarpomanova, B. Rizov, T. Dimitrova & H. Kukova, “Bulgarian Sense-Annotated Corpus—Results and Achievements”, *FASSBL7*, 2010: 41.
8. Y. Wu, P. Jin, Y. Zhang & S. Yu, “A Chinese corpus with word sense annotation” Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead, *Springer*, Berlin Heidelberg, 2006.

References

9. S, Urooj, S, Hussain, F, Adeeba, F. Jabeen and R. Perveen, "CLE Urdu Digest Corpus", in the Proc. of *Conference on Language and Technology 2012* (CLT12), Lahore, Pakistan, 2012.
10. A. Zafar, A. Mahmood, F. Abdullah, S. Zahid, S. Hussain, and A. Mustafa, "Developing Urdu WordNet Using the Merge Approach ", in the Proceedings of *Conference on Language and Technology 2012* (CLT12), Lahore, Pakistan.
11. Hussain, S., "*Finite-State Morphological Analyzer for Urdu*", National University of Computer and Emerging Sciences, (2004), Lahore, Pakistan.
12. M. Butt, "The light verb jungle: Still Hacking Away", in Workshop on Multi-Verb Constructions, 2003.