Semantic Tagging, Multilinguality, Development and Applications

Scott Piao

School of Computing and Communications
Lancaster University
Lancaster
UK
Email: s.piao@lancaster.ac.uk

UCREL Summer School in Corpus-based NLP 2017

Outline of This Course

- Brief introduction to Lancaster UCREL USAS framework.
- Introduction to development of multilingual semantic taggers and HTST.
- Applications of the semantic tagger.
- Accessing the taggers as a web service using programs.
- Practical exercise of tagging multilingual texts (or your own texts) and extracting various semantic information based on provided sample code - *you can also try anything you are interested*.

Brief History of USAS Semantic Tagger

- USAS Semantic tagger has been developed at UCREL, Lancaster University over the past two decades (Rayson et al., 2004) -- a core engine for the web-based corpus analysis website Wmatrix (http://ucrel.lancs.ac.uk/wmatrix/)
- Meanwhile, initially developed for English, the USAS semantic tagger has also been ported for other languages through some projects and in-house work, and a Java version was developed for handling multilingual data.
- Recently, the semantic tagger has been expanded to tag English text in a fine-grained semantic categories using a large English thesaurus, leading to HTST tagger (Piao et al., 2017).
- Semantic tagger software have been developed for eleven languages – taggers for nine languages publicly accessible.
- For further details about USAS, see website http://ucrel.lancs.ac.uk/usas/.

USAS Semantic Annotation Tagset

--- 22 Major categories and 232 sub-categories

*(In teaching materials, see file "~/**tag-sets/USAS-Semantic-Tagset.pdf**")*

A General and abstract terms	**B** The body and the individual	**C** Arts and crafts	**E** Emotion
F Food and farming	**G** Government and public	**H** Architecture, housing and the home	**I** Money and commerce in industry
K Entertainment, sports and games	**L** Life and living things	**M** Movement, location, travel and transport	**N** Numbers and measurement
O Substances, materials, objects and equipment	**P** Education	**Q** Language and communication	**S** Social actions, states and processes
T Time	**W** World and environment	**X** Psychological actions, states and processes	**Y** Science and technology
Z Names and grammar			

Course-grained Generic Semantic Classification

- Based on Tom McArthur's Longman Lexicon of Contemporary English (McArthur, 1981), the USAS tagset provides a coarsely-grained lexical semantic classification scheme.
- It is a generic scheme, not constrained to specific domain/s.
- Can be used to analyse high level abstract semantic information of text, such as key topics of documents and semantic linking between documents.
- Provide extra codes to denote information such as positive/negative, gender etc.
  - Example of tags:
    - ***E4.1+*** and ***E4.1-*** denotes ***happiness*** and ***sadness***;
    - ***S4f*** and ***S4m*** indicate ***female*** and ***male relatives***;
    - Etc.
- For detailed explanation about USAS semantic categories and tags, see document "***~/tag-sets/usas_tagset-explanation.pdf***" included in the teaching materials.

USAS Framework
(Java version)

- Semantic lexicon resources
  - Single word dictionary
    - bank NN1 I1/H1 I1.1/I2.1c W3/M4 A9+/H1 O2 M6
  - Multi-word expression dictionary, including templates.
    - giv*_* {R*/Np/PP*} away_* A9- A10+ S4
- Template rules
  - {kind}[A4.1] *_IO *_N*
- Context rules
- Achieves 91% accuracy on modern English
- For further details, see paper:
  - Rayson, Paul, Dawn Archer, Scott Piao, Tony McEnery (2004). The UCREL semantic analysis system. In proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks, LREC 2004, Lisbon, Portugal, pp. 7-12.

Sample of Single Word Lexicon

Manchester	NP1	Z2 Z3
Mancunian	JJ	Z2 Z2/Q3
Mancunian	NN1	Z2/S2mf Z2/Q3
Mandarin-speaking	JJ	Z2/Q3
Mandela	NP1	Z1mf
Mandella	NP1	Z1mf
Manderville	NP1	Z2
Mandeville	NP1	Z2
Mandy	NP1	Z1f
…		
man-to-man	JJ	S5- S1.2.1+ A5.2+ A5.4+
manacles	NN2	O2
manage	VV0	S7.1+ A1.1.1 X9.2+
manageable	JJ	A12+
managed	JJ	S7.1+ A1.1.1 X9.2+
management	NN	S7.1+
management-style	JJ	S7.1+
manager	NN1	S7.1+/S2mf K1/S7.1+/S2mf K5/S7.1+/S2mf
manageress	NN1	S7.1+/S2.1f
manageress	VV0	S7.1+
managerial	JJ	S7.1+

Sample of Multi-Word Expression (MWE) Lexicon

at_II the_AT very_RG least_DAT	A13.7
at_II the_AT very_RG minimum_*	A13.7
at_II the_AT {J*/UH} offset_NN1	T2+
at_II the_AT {J*} forefront_NN1 of_IO	A11.1+
at_II the_AT {J*} mercy_NN1 of_IO	S7.1-
at_II the_AT {J*} moment_NN1	T1.1.2
at_II the_AT {J*} outset_NN1	T2+

Context Rules

Context rules help disambiguate word senses subject to strong context patterns, as shown below.

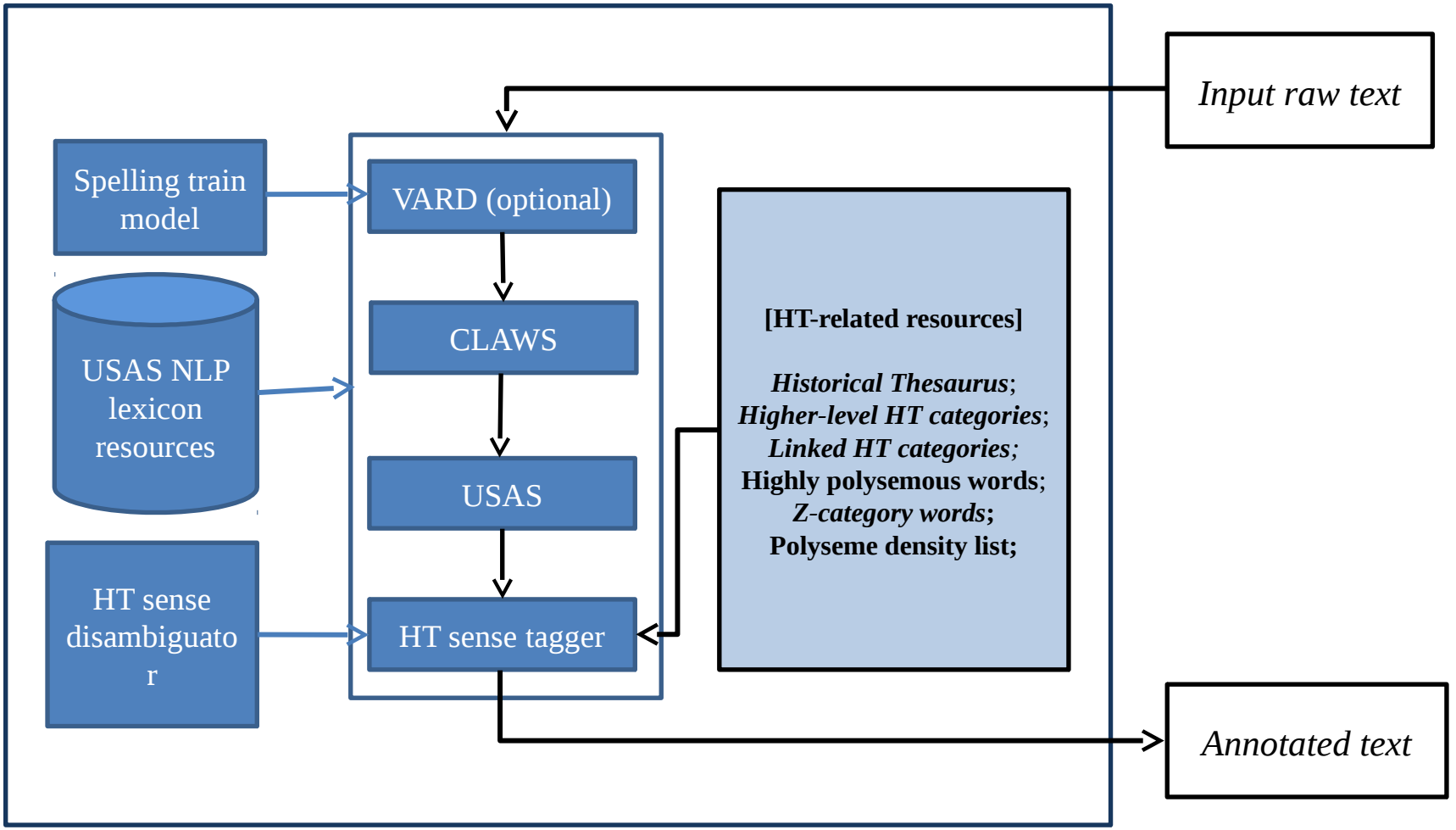
Disambiguation rules for word "work":
*_II {APPGE} {work}[I3.1]
*_II *_NN* *_GE {work}[I3.1]
_NN *_VB* {at} {work}[A1.1.1]
*_PN1 *_VB* {at} {work}[A1.1.1]

Where,
I3.1 *denotes "Work and employment: Generally"*
A1.1.1 *denotes "General actions, making etc."*

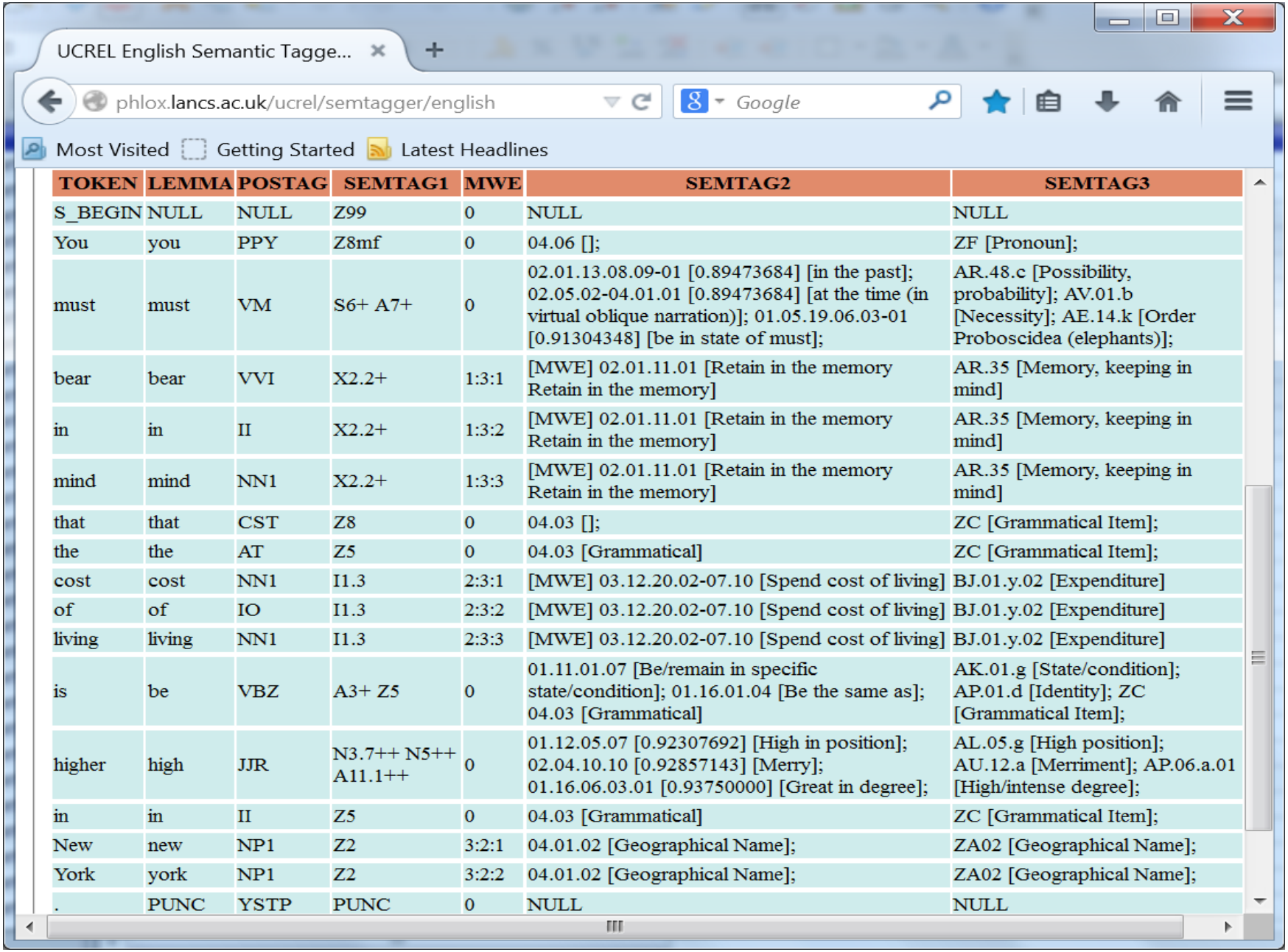
Extension of English Semantic Tagger: HTST tagger

- Recently, the USAS was extended to tag English text in a finer-grained semantic classification scheme based on a large-scale English Historical Thesaurus in the Samuels Project, named HTST (Piao et al., 2017; Alexander et al., 2015).
- HTST employs 4,033 semantic categories, derived from the English Historical Thesaurus developed by Glasgow University and Oxford University Press (*http://historicalthesaurus.arts.gla.ac.uk*).
- Employs context-based disambiguation methods at word and annotation levels to disambiguate words' semantic categories.
- Employs time filter for improving tagging accuracy, particularly useful for tagging historical text.
- For definitions of the HTST tags, see document “~/tag-sets/htst-samuels-tagset.xlsx”

Architecture of Historical Thesaurus-based Semantic Tagger (HTST)



HTST Sample Output



TOKEN	LEMMA	POSTAG	SEMTAG1	MWE	SEMTAG2	SEMTAG3
S_BEGIN	NULL	NULL	Z99	0	NULL	NULL
You	you	PPY	Z8mf	0	04.06 [];	ZF [Pronoun];
must	must	VM	S6+ A7+	0	02.01.13.08.09-01 [0.89473684] [in the past]; 02.05.02-04.01.01 [0.89473684] [at the time (in virtual oblique narration)]; 01.05.19.06.03-01 [0.91304348] [be in state of must];	AR.48.c [Possibility, probability]; AV.01.b [Necessity]; AE.14.k [Order Proboscidea (elephants)];
bear	bear	VVI	X2.2+	1:3:1	[MWE] 02.01.11.01 [Retain in the memory Retain in the memory]	AR.35 [Memory, keeping in mind]
in	in	II	X2.2+	1:3:2	[MWE] 02.01.11.01 [Retain in the memory Retain in the memory]	AR.35 [Memory, keeping in mind]
mind	mind	NN1	X2.2+	1:3:3	[MWE] 02.01.11.01 [Retain in the memory Retain in the memory]	AR.35 [Memory, keeping in mind]
that	that	CST	Z8	0	04.03 [];	ZC [Grammatical Item];
the	the	AT	Z5	0	04.03 [Grammatical]	ZC [Grammatical Item];
cost	cost	NN1	I1.3	2:3:1	[MWE] 03.12.20.02-07.10 [Spend cost of living]	BJ.01.y.02 [Expenditure]
of	of	IO	I1.3	2:3:2	[MWE] 03.12.20.02-07.10 [Spend cost of living]	BJ.01.y.02 [Expenditure]
living	living	NN1	I1.3	2:3:3	[MWE] 03.12.20.02-07.10 [Spend cost of living]	BJ.01.y.02 [Expenditure]
is	be	VBZ	A3+ Z5	0	01.11.01.07 [Be/remain in specific state/condition]; 01.16.01.04 [Be the same as]; 04.03 [Grammatical]	AK.01.g [State/condition]; AP.01.d [Identity]; ZC [Grammatical Item];
higher	high	JJR	N3.7++ N5++ A11.1++	0	01.12.05.07 [0.92307692] [High in position]; 02.04.10.10 [0.92857143] [Merry]; 01.16.06.03.01 [0.93750000] [Great in degree];	AL.05.g [High position]; AU.12.a [Merriment]; AP.06.a.01 [High/intense degree];
in	in	II	Z5	0	04.03 [Grammatical]	ZC [Grammatical Item];
New	new	NP1	Z2	3:2:1	04.01.02 [Geographical Name];	ZA02 [Geographical Name];
York	york	NP1	Z2	3:2:2	04.01.02 [Geographical Name];	ZA02 [Geographical Name];
.	PUNC	YSTP	PUNC	0	NULL	NULL

Multilingality of Semantic Tagging

- Multilinguality is an important aspect of NLP, and so to semantic analysis.
- Would be nice to create an NLP ecosystem for multilingual semantic tagging and analysis under the same semantic classification framework.
- For this purpose, the USAS tagger has been continuously extended to cover more and more languages.
- The first step is to construct semantic lexicons for new languages.
- Currently, the USAS lexicons cover fourteen languages, including Italian, Portuguese, Chinese, Spanish, Arabic, Russian, French, Czech, Finnish, Dutch, Malaysian, Welsh, Swedish and Urdu besides English.
- Based on the lexicons, semantic tagging software have been developed for Finnish, Russian, Italian, Portuguese, Dutch, Chinese, Spanish, French, Swedish and Welsh (Russian and Finnish taggers are not publicly available).
- The same software framework is used for different languages with minor adjustments.
- Semantic taggers are in different stages of development for different languages, hence they provide various lexical coverages and accuracies for different languages.

Multilingual Semantic Lexicon Construction

- A critical part of multilingual semantic tagger development is to construct semantic lexicons for the languages.
- Various approaches have been employed so far:
  - Automatically translating the core English semantic lexicon using bilingual dictionaries and other publicly available lexicons.
  - Using crowd-sourcing methods to clean and expand the automatically generated lexicons.
  - Where possible, using bilingual parallel corpora to align words across languages, thereby allowing the application of above two methods.
  - Using machine translation tools to directly translate existing lexicons into new languages.
  - Manually cleaning and curating the lexicons whenever possible.
  - *There should be more good methods … that we can try.*

CorCenCC Project for Welsh Language

- Corpus and tool development for new language is an important research theme and gains support.
- CorCenCC (The National Corpus of Contemporary Welsh) Project is funded by UK ESRC and AHRC with £1.8 million for building Welsh corpus and NLP tools (see project website: http://www.corcencc.org/).
- UCREL team has been working on the development of a Welsh semantic tagger in this project.
  - A large semantic lexicon has been built.
  - A prototype Welsh semantic tagger has been developed, see demo website: http://phlox.lancs.ac.uk/ucrel/semtagger/welsh
  - Welsh semantic tagger is integrated into UCREL's multilingual tagger package.
- If you are interested in Welsh language, you are welcome to collaborate with us.

Current USAS Lexicons for 14 Languages (excluding English)

Language	Single Word Entries	MWE Entries	Tagger Created?
Arabic	31,154	0	N
Chinese	64,541	19,048	Y
Czech	28,161	0	N
Dutch	4,220	0	Y
Finnish	46,225	4,422	*Y
French	2,754	0	Y
Italian	13,098	5,622	Y
Portuguese	13,499	1,781	Y
Russian	17,443	713	*Y
Spanish	9,710	4,840	Y
Urdu	1,765	235	N
Welsh	143,280	sample	Y
Swedish	18,080	0	Y
Malay	64,863	0	N

Lexical Coverage Estimation for 13 Languages

No	Language	Blogs (million words)	News (million words)	Average	Tagger or Lexicon only?
1	Finnish	95.98	95.89	95.93	Tagger
2	Italian	91.14	89.34	90.24	Tagger
3	Czech	87.95	86.05	86.99	Tagger
4	Russian	84.93	86.66	85.79	Tagger
5	Chinese	82.98	79.36	81.17	Tagger
6	Portuguese (EU)	76.79	77.47	77.13	Tagger
7	Portuguese (BR)	76.11	77.75	76.93	Tagger
8	Dutch	61.55	59.87	60.71	Tagger
9	Spanish (EU)	57.81	55.73	56.77	Tagger
10	Spanish (SA)	57.20	56.11	56.65	Tagger
11	Arabic	86.43	91.33	88.88	Lexicon only
12	Urdu	86.26	84.21	85.24	Lexicon only
13	Malay	53.83	54.91	54.37	Lexicon only

Need for Expansion and Refining of Multilingual Lexicons

- We need to continuously expand the multilingual lexicons, both lexical coverage and number of languages.
- We need to improve the quality of the semantic lexicons by clearing errors and refining semantic classification of the lexicon entries.
- Multiword Expression (MWE) lexicon, including templates, is a powerful way of improving the semantic tagging, we need to search for efficient ways to extracting high quality MWE lexicons and their semantic categories.
- Everyone is welcome to participate this research!

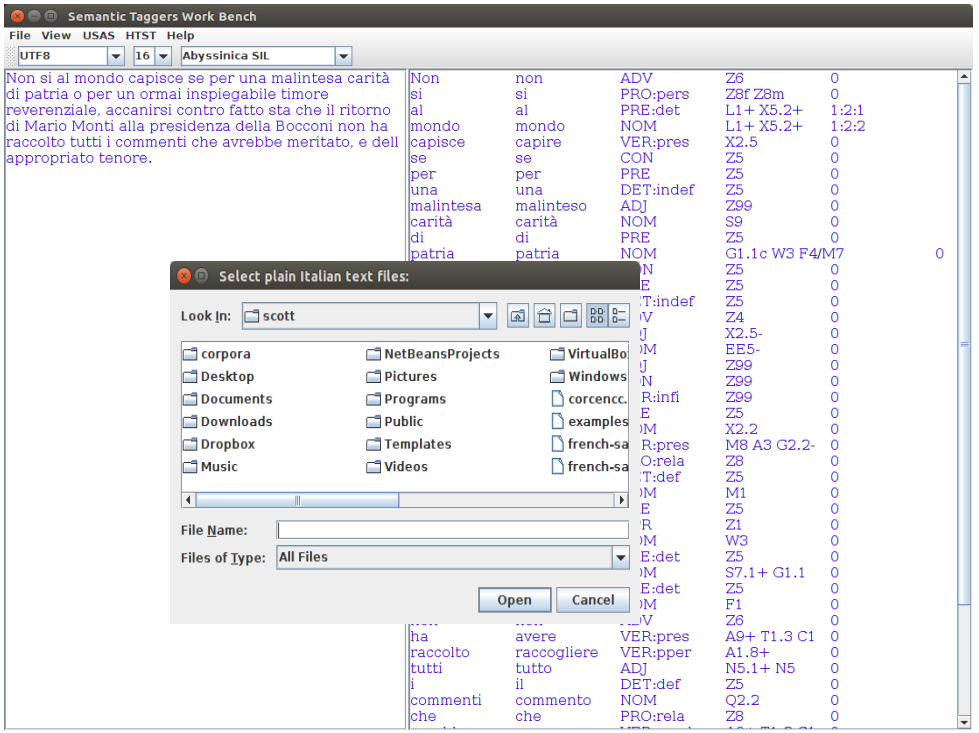
(***Current versions of the multilingual lexicons are available at website: https://github.com/UCREL/Multilingual-USAS***)

Multilingual Semantic Taggers

- The main software framework is built in Java.
- POS taggers include:
  - CLAWS for English;
  - Stanford POS Tagger for Chinese;
  - TreeTagger for Italian, Spanish, Portuguese, French, Dutch, Swedish;
  - WNLT (https://sourceforge.net/projects/wnlt) for Welsh (a new Welsh POS tagger is under development in CorConCC Project);
  - Grampal tagger for Spanish (http://cartago.lllf.uam.es/grampal/grampal.cgi).
- The semantic taggers are built into web services.
- Web URL for demo site:
  - http://phlox.lancs.ac.uk/ucrel/semtagger/{language}
  - Language options: english, italian, spanish, chinese, portuguese, french, dutch, welsh, swedish.
- A desktop application (user interface) has been developed for easily accessing and processing text (see zipped file “sem-muling-tagger-gui.zip” or “sem-muling-tagger-gui.tar.gz” in the teaching materials).

Two Graphical User Interfaces of Multilingual USAS Services

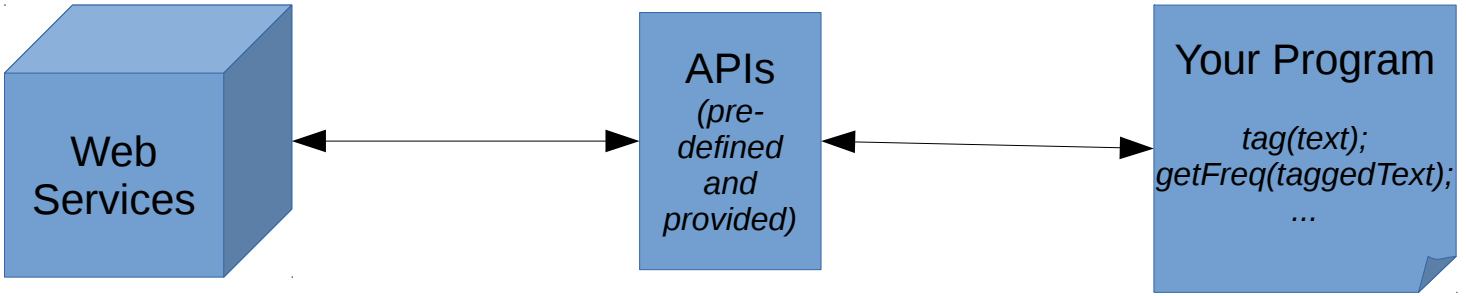
Web demo interface



Desktop GUI application

Programmatically Accessing the Semantic Taggers via Web APIs

- Now let's learn how to use the taggers within your own computer programs.
- Currently the web services of semantic taggers provide API access.
- API (Application Programming Interface) is a set of clearly defined methods for communicating between software, web services in our case.
- Web service API access allows convenient remote access to software functionalities.



UCREL Semantic Tagger Web APIs

- Currently the UCREL semantic tagger web services provide the following APIs for calling the taggers of nine languages:
  - String *SemanticTaggerClient*.*tagChiText*(String); – Chinese
  - String *SemanticTaggerClient*.*tagDutText*(String); – Dutch
  - String *SemanticTaggerClient*.*tagEngText*(String); – English
  - String *SemanticTaggerClient*.*tagFrenText*(String); – French
  - String *SemanticTaggerClient*.*tagItaText*(String); – Italian
  - String *SemanticTaggerClient*.*tagPortText*(String); – Portuguese
  - String *SemanticTaggerClient*.*tagSpanText*(String); – Spanish
  - String *SemanticTaggerClient*.*tagSwedText*(String); – Swedish (TreeTagger)
  - String *SemanticTaggerClient*.tagSpanTextGrampal(String); - Spanish (Grampal)
  - String *WelshTaggerClient*.*semTagWnlt*(String); – Welsh*

*Another version of Welsh tagger is under development in CorCenCC Project.

Semantic Taggers' Output Format

- For English, the semantic tagger's output is in seven columns separated by TAB:
  - *TOKEN LEMMA POSTAG USAS-TAG MWE_CODE HT-TAG*
- For other languages, the output contains five columns separated by TAB:
  - *TOKEN LEMMA POSTAG SEMTAG MWE_CODE*
- For this session, *TOKEN*, *LEMMA* and USAS-TAG/HT-TAG provide the most useful information.
- Three web sources for checking the meaning of semantic tags.
  - For USAS-TAG definitions, see http://ucrel.lancs.ac.uk/usas/USASSemanticTagset.pdf
  - For detailed explanations about USAS-TAGs, see http://ucrel.lancs.ac.uk/usas/usas_guide.pdf
  - For HT-TAG, see

    http://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/samuels/ -> Project outputs -> Thematic Categories.
  - These docs are also included in the teaching materials.

A Sample SEM-Tagged Output

#WORD	LEM	POS	SEM-TAG	MWE-FLAG
Desde	desde	PREP	Z5 N3.3 T1	0
la	el	ART,fem,sing	Z5	0
1	1	Q	N1	0
hasta	hasta	PREP	Z5 T1 N3.3	0
las	el	ART,fem,plu	Z5	0
10	10	Q	N1	0
estará	estar	V,sing,3,fut_ind	A3+ M8	0
cerrado	cerrado	ADJ,masc,sing	A10- T2-	0
.	.	PUNCT	PUNCT	0

*English translation: “From 1 to 10 it will be closed.”

How to Call Tagger Service

- With the APIs (library programs are provided in teaching materials), you can tag Italian and Welsh texts as follows:

  - Create tagger object:
    SemanticTaggerClient semanticTagger = new SemanticTaggerClient();
    WelshTaggerClient welshTagger = new WelshTaggerClient();
  - Then call tagging functions for each language:
    String taggedItalianText = semanticTagger.tagItaTex(italianText);
    String taggedWelshText = welshTagger.semTagWnlt(welshText);
  - *Then do something interesting to you with the tagged texts, e.g.*
    checkSemanticLinksBetween(taggedItalianText, taggedWelshText);

- *To avoid overloading server computer, limit size of each input text to 10,000 words/tokens for now. If process bigger texts, break them down into smaller sections.*

A Sample Program

```
import ac.uk.lancs.ucrel.semtaggers.web.clients.SemanticTaggerClient;
import semtagger.welsh.webservice.client.WelshTaggerClient;

public class UcrelSemTaggerTester {

    public static void main(String[] args) {

        //Create tagger object “semanticTagger”
        SemanticTaggerClient semanticTagger = new SemanticTaggerClient();

        //Prepare a text
        String text = "I enjoy this summer school.";

        //Call sem-tagging function “tagEngText()” of “semanticTagger” to tag the above text.
        String taggedText = semanticTagger.tagEngText(text);

        //Save the tagged text in a file:
        //First create a file managing object from the provided class “OpenSaveFile”
        OpenSaveFile fileOpenSave = new OpenSaveFile();

        //Call function “saveTextToFile()” to save the tagged text in a file with a specified encoding.
        fileOpenSave.saveTextToFile(taggedText, “sem-tagged-text.txt”, “UTF8”);

        //Continue to do something you like with the tagged text

    }
}
```

Provided Teaching Materials

(*Check SS2017 teaching materials folder:* ***S3-semtagging-multiling-applications***)

- Some programs and sample code are provided in the teaching materials site to help you get on with the programming:
  - The desktop GUI package for testing, see folder "~/sem-muling-tagger-gui".
  - Library Jar files for calling the tagger web service to be included in classpath, see folder:

    "~/sample-programs-4-tagging-text/lib/"
  - Sample code for accessing the tagger service, see folder "~/tagger-service-access-sample-code":

    a) "UcrelMultilingSemTaggerTester.java" provides a sample code for using the client program,

    b) "TagTextsInFile.java" for tagging texts stored in files.

    c) "SemTagFreqExtractor.java" for extracting frequencies of semantic tags.
  - Folder "~/sample-texts" and "files-4-test" contain sample texts of nine languages for testing the semantic taggers. Of course you can use your own texts.
  - All sample texts are encoded in "UTF8", and the semantic tagger uses this encoding by default.
- Check the teaching materials site for possible updates.

Useful Tips for Running Java Programs

- Compile basic Java program:

  >javac {program-name.java}

- To include library – often JAR files
  - *javac -classpath {paths-to-library-files} {program-name.java}, e.g.*

    >javac -classpath .:lib/* SemTagFreqExtractor.java

- If computer complains about non-ASCII code, then

  >javac -encoding {encoding-name} -classpath {path-to-library-file} {program-name.java}

  - *Note: For encoding, "UTF8" is suggested for multi-language text processing.*
- To run a compiled Java program, including a library, use a command like

  > java -cp .:lib/* SemTagFreqExtractor

- *If you are using IDE tool, you can include the library JAR files in project's classpath.*

Applications

- The semantic taggers can be applied in various tasks, e.g.
  - Identify main semantic themes of documents, e.g. find most frequent semantic categories.
  - Link multilingual documents which share specific semantic themes.
  - Extract key semantic patterns of documents and link such documents across languages.
  - Information extraction, knowledge extraction.
  - Improve information retrieval.
  - ...
  - You may know more interesting applications – try it out during the practical session.

Suggested Tasks for Practical Session

- Collect some texts written in English and another language (that you are interested in) from web pages or from your own archive, keep them in separate files.
- By modifying the provided sample programs, tag the texts and save them into separate output files.
- Further update the sample program to check the frequencies of the USAS semantic tags in each tagged corpus and find five most frequent semantic tags (excluding grammatical words) – treating them as themes of the texts.
- Further update the program to check if any English texts share any themes with non-English texts. If you find any matched texts of different languages, print out the theme semantic tags and file names linked by them in a tabular format, e.g.

  file1, file2, shared-semantic-tags
  ...
- *Of course, you can do more complex and interesting things if you want!*

Related Papers

- Piao, Scott, Fraser Dallachy, Alistair Baron, Jane Demmen, Steve Wattam, Philip Durkin, James McCracken, Paul Rayson, Marc Alexander (2017). Time-Sensitive Historical Thesaurus-based Semantic Tagger for Deep Semantic Annotation . Computer Speech and Language, vol. 46, Elsevier. pp. 113-135. doi:10.1016/j.csl.2017.04.010.
- Piao, Scott, Paul Rayson, Dawn Knight, Gareth Watkins and Kevin Donnelly (2017). Towards a Welsh Semantic Tagger: Creating Lexicons for A Resource Poor Language. The Corpus Linguistics 2017 Conference, 24-28 July 2017 at University of Birmingham, Birmingham, UK.
- Piao, Scott, Paul Rayson, Dawn Archer, Francesca Bianchi, Carmen Dayrell, Mahmoud El-Haj, Ricardo-María Jiménez, Dawn Knight, Michal Křen, Laura Löfberg, Rao Muhammad Adeel Nawab, Jawad Shafi, Phoey Lee Teh, Olga Mudraya (2016). Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages. In Proceedings of The 10th Edition of the Language Resources and Evaluation Conference (LREC2016), held during 23-28 May 2016 in Portorož, Slovenia.
- Piao, Scott, Francesca Bianchi, Carmen Dayrell, Angela D'Egidio and Paul Rayson (2015). Development of the Multilingual Semantic Annotation System. The 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015), Denver, Colorado, USA.
- Alexander, Marc, Fraser Dallachy, Scott Piao, Alistair Baron, Paul Rayson (2015). Metaphor, Popular Science and Semantic Tagging: Distant reading with the Historical Thesaurus of English. Digital Scholarship in the Humanities, Oxford University Press, UK.
- Rayson, Paul, Dawn Archer, Scott Piao and Tony McEnery (2004). The UCREL semantic analysis system. In Proceedings of LREC-04 Workshop: Beyond Named Entity Recognition Semantic Labeling for NLP Tasks, pp. 7-12. Lisbon, Portugal.
- McArthur, Tom (1981). Longman Lexicon of Contemporary English. Longman London.