CSE6390 3.0 Special Topics in AI & Interactive Systems II
Introduction to Computational Linguistics

Insructor: Nick Cercone – 3050 CSEB – nick@cse.yorku.ca
Tuesdays,Thursdays 10:00-11:30 – South Ross 104
Fall Semester, 2010

Resources

Go to the following website for the most comprehensive listing of CL resources (slightly edited version is shown below)

http://users.cs.dal.ca/~vlado/nlp/

General Links

- Central Resources
  - ACL Anthology - A Digital Archive of Research Papers in Computational Linguistics
- Associations
  - ACL
  - EACL - ACL European chapter
- Journals
  - ACM Transactions on Speech and Langauge Processing
- Institutions and Groups
  - DNLP - Dalhousie Natural Language Processing Group
  - CSLI, Stanford University
  - Computational Linguistics research group at University of Toronto
  - Illinois Natural Language Research - NLP group at the University of Illinois at Urbana Champaign
  - NLP group at Technical University of Catalonia, Spain
  - ISSCO - Language group at University of Geneva
  - The British Computer Society Natural Language Translation Specialist Group
  - Tsujii Laboratory, Japan
  - NLG - Natural Language Group at UCS/ISI
  - The Center for Language and Speech Processing at the Johns Hopkins University
  - SIGGEN - ACL Special Interest Group on Generation - (Natural Language Generation)
  - Research Group in Computational Linguistics at University of Wolverhampton
- Conference links
  - NLP related conferences - collected by Vlado Keselj (conferences.txt - human/machine-friendly source file)
  - Evangelos Milios's page (contains conference links)
  - DNLP research group's web page
  - Fuchun Peng's list of conferences
  - Conferences at ACL NLP/CL Universe, by Dragomir Radev
- E-mail lists
  - Corpora List
  - HPSG mailing list
  - The LFG mailing list
  - List archives at listserv.linguistlist.org
- NLP Courses
  - http://aclweb.org/aclwiki/index.php?title=List_of_NLP/CL_courses
- General Human Language Resources
  - Ethnologue - Languages of the World - "An encyclopedic reference work cataloging all of the world's 6,912 known living languages"
- Other General Resources
  - BioNLP Resources - by Alex Morgan
  - The ACL NLP/CL Universe
  - Christopher Manning's list of resources
  - HPSG Bibliography
  - Linguistlist.org
  - Software Tools for NLP
  - Transformation-Based Learning Bibliography
  - Resource list by Wirote Aroonmanakun [Link1]

Speech processing

- CMU Sphinx - The CMU Sphinx Group Open Source Speech Recognition Engines
- eSpeak - text to speech open source software
- The Festival Speech Synthesis System - by the Centre for Speech Technology Research, Univ.of Edinburgh

- Festival at CMU
- PRAAT: Phonetics and Speech Tools
- The EMU Speech Database System
- Example of prosodic annotation data
- ToBI Annotation document
- Phonology

    Merriam-Webster's Pronunciation Guide

    Merriam-Webster's Pronunciation Symbols

    Automatic Phonetic Transcription Tools

        A summary on corpora list (2003)

        t2p: Text-to-Phoneme Converter Builder - in Perl by Kevin Lenzo

- Commercial

    ScanSoft

    Nuance Communications Inc.

N-gram analysis

- Text::Ngrams Perl Package - Flexible Ngram analysis (for characters, words, and more); on CPAN; by Vlado Keselj
- Ngram Statistics Package in Perl, by T. Pedersen at al.
- Text::Ngram Perl Package by Simon Cozens
- Perl script ngram.pl by Jarkko Hietaniemi
- Waterloo Statistical N-Gram Language Modeling Toolkit - in C++ by Fuchun Peng
- Suffix Arrays for Ngrams

    Suffix Arrays - description and implementation by Douglas McIlroy, with implementation by Sean Quinlan and Sean Dorward

Preprocessing

- Transcription and Encoding Schemes

    Thai transcription Web service [Link1]

- Sentence Splitters (sentencizers, sentence boundary detectors)

    Perl5 HTML-Summary module (CPAN)

    Perl Lingua::EN::Sentence module (CPAN)

    A Java splitter

    In C (source code included)

    Adwait Ratnaparkhi's MXTERMINATOR in Java

    LTG TTT system

    cogcomp

    zzheng: on-line splitter, CGI script

    Guenther: CGI script, to appear

- Word segmentation

    Thai word segmentation, Web service [Link1]

- Stop-word Removal

    A list of stop words at http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words

Morphology

- Letter-to-Phoneme Pascal Challenge 2006
- Unsupervised segmentation of words into morphemes -- Challenge 2005
- Using eigenvectors of the bigram graph to infer morpheme identity - by Mikhail Belkin and John Goldsmith, 2002
- Stemmers and Lematizers

    The Porter Stemming Algorithm ('official' page) - maintained by Martin Porter, many programming languages. This is the official Porter stemmer. Have in mind that many modified implementations are around, some giving incorrect stems for 14% words (e.g. try 'agreement' - it is supposed to be stemmed to 'agreement' and not 'agreem').

    Snowball - 'small string processing language designed for creating stemming algorithms', by Martin Porter, includes stemmers for several languages

    Lovins stemmer - The (un)official Lovins stemmer page

    Porter stemmer implemented in C, Pascal, Visual Basic, and Java

Finite State Methods

- intex - Linguistic Development Environment by Max Silberztein
- UNITEX - Corpus processing sytem, GPL licence, implemented in C/C++ and Java; spawned from the same project as intex and NooJ
- Nooj - Linguistic development environment - Related to intex, by Max Silberztein.
- JFLAP - Java Formal Langauge and Automata Package

POS Tagging

- Software Plaza:Brill's tagger
- AI repository: Brill's tagger
- TnT -- Statistical Part-of-Speech Tagging
- QTag - a probabilistic POS tagger - language independent, implemented in Java, by Oliver Mason (c) 1994-2003
- Ingo's collection of POS taggers
- CLAWS part-of-speech tagger
- UCREL CLAWS7 Tagset
- AI repository: taggers

- Brazilian Portuguese POS Tagger (based on QTAG)
- Lingua::EN::Tagger - POS tagger for English; Perl module; uses HMM bigram word model; by Maciej Ceglowski and Aaron Coburn
- POS tagger in Perl/Tk by Kristie Seymore
- BNC POS tags
- A Practical Part-of-Speech Tagger (1992) - Cutting et al.; Xerox code, Application to Spanish
- SVMTool - Open source POS tagger based on Support Vector Machine
- Stanford Log-linear Tagger

Document Clustering

- Tutorial on clustering Large and High-Dimensional data by Nicholas et al. - On CIKM 2003
- Clustering and Segmentation software on KDnuggets
- CLUTO - Software Package for Clustering High-Dimensional Datasets
- Matlab Clustering Package by Frank Dellaert

Terminology Extraction

- Gensen Web - An automatic domain terminology extraction system

Text Categorization (TC)

- Spam detection and E-mail classification

TREC 2005 - Spam Track preliminary guidelines

Papers from CEAS 2005 (Conference on Email and Anti-Spam)

TREC 2005 Spam Filter Evaluation Tool Kit

TREC 2005 Corpus (92,000 messages - 42,000 ham; 50,000 spam)

TREC Spam Evaluation 2005

Gord Cormack's paper

Paul Graham's list of useful links

2005 MIT spam conference

2005 conference on email and spam (CEAS)

A study in analyzing spam test results

SpamAssasin

SpamAssasin Perl package

POPFile - Open-source e-mail classification software, in Perl

SpamBayes [Link1]

Ion Androutsopoulos's Publications

lingspam_public.tar.gz data

- Encoding identification

Mozilla charset detectors [Link1]

jchardet, Java port of Mozilla's automatic charset detection algorithm

Google on encoding identification

- Language identification

Note: The following references are relevant written language identification. The spoken language identification is a different area of research and related references are not included here.

Language identification tools, by Gertjan van Noord (TextCat)

On-line tool by Steve Huffman

Chapter on Automatic Language Identification - in Survey of the State of the Art in Human Language Technology by several editors

A Language identification tool at Fagan finder

Another language identification tool

Language identifier by Ken Beesley

DRUID, a language identification tool

Specifying language excerpts in XML

SILC project at RALI

Language Identification tool - by Veristage; minimum 40 characters

Language identification and IT: Addressing problems of linguistic diversity on a global scale - by Peter Constable and Gary Simons, SIL International; about language tagging

Language identification flashcard - by US Dept. of Commerce

Comment by J. Goodman on a Physics paper about Language Trees and Zipping, which got a lot of press coverage in 2001

Universal Declaration of Human Rights - UN, in 363 languages (17 Jun 2004)

- Sentiment classification

Movie Review Data - Corpus for sentiment classification by Bo Pang

Thumbs up? Sentiment Classification using Machine Learning Techniques - by Bo Pang, Lilian Lee, and Shivakukmar Vaithyanathan, EMNLP-2002

- Authorship attribution and Plagiarism detection (AATT)

Ad-hoc Authorship Attribution Competition, Patrick Juola, 2004

- Topic categorization

Reuters21578 Collection (used for categorization task)

- Other

W. J. Teahan. Text Classification and Segmentation Using Minimum Cross-Entropy. In Proceedings of the

International Conference on Content-based Multimedia Information Access (RIAO 2000), pages 943-961. C.I.D.-C.A.S.I.S, Paris, France, 2000. ISBN 2-905450-07-X.
http://www.scms.rgu.ac.uk/staff/smc/researchcoord/staff_publications/wjt.html

Text Summarization
- Text Summarization site - by Dragomir Radev
- "Statistics-Based Summarization --- Step One: Sentence Compression," (K. Knight and D. Marcu), National Conference on Artificial Intelligence (AAAI), 2000.

Dictionaries and Lexicons
- ACL-SIGLEX - ACL Special Interest Group on Lexicon
- Dictionary development
  TEI (Text Encoding Initiaive): 12 Print Dictionaries
  TEI-L e-mail list
  The DICT Development Group
  dict-compare - Perl script
- On-line dictionaries
  YourDictionary.com - likely the most comprehensive index of dictionaries available on the web, collected by Robert Beard
  Recnik.com - Serbian-English-Serbian bidirectional on-line dictionary
  Rjecnik.com - Croatian-English-Croatian bidirectional on-line dictionary
  Ba.Rjecnik.com - Bosnian-English-Bosnian bidirectional on-line dictionary
  http://www.language-archives.org/
  http://www.worldlanguage.com/ProductTypes/Dictionary.htm
  http://www.dictionarium.com/
  http://www.lai.com/glossaries.html
  http://www.wordgumbo.com/index.htm
  http://www.yourdictionary.com/

Lexical Semantics
- WordNet
  WordNet home
  WordNet On-line
  Global WordNet Association
  MultiWordNet project (Italian)
  WordNet::Similarity - Perl module implementing several lexical semantic similarity measures by Siddharth Patwardhan and Ted Pedersen
  Java WordNet library, and some other interesting NLP software [Link1]
- Word Sense Disambiguation (WSD)
  Evaluation exercises for Word Sense Disambiguation, organized by ACL-SIGLEX
  Senseval-2 System Code and Documentation, by Ted Pedersen
  Senseval-2 Software and data, by Saif Mohammad

Unification
- Theory
  "Unification: A Multidisciplinary Survey," Kevin Knight, ACM Computing Surveys, 21(1), 93-124, 1989.
- Practice: Unification-based Systems
  ALE unification-based,parser - coverage: medium
  LKB unification-based,parser - coverage: medium
  PC-PATR unification-based,parser - coverage: small
  Stefy unification-based,parser - coverage: small
  Utool 3.0; The Swiss Army Knife of Underspecification - in Java; several input formats including LKB

Grammar Formalisms
- Unification-based grammars
- Head-driven Phrase Structure Grammar (HPSG)
  HPSG - Stanford University
  HPSG - Ohio State
  HPSG Bibliography
  Slavic languages in HPSG, Warsaw
  Proceedings of HPSG 2004
  A Survey of Systems for Implementing HPSG Grammars - by Leonard Bolc, Krzysztof Czuba, Anna Kupsc, Malgorzata Marciniak, Agnieszka Mykowiecka and Adam Przepirkowski, Technical report, 96
- Lexical Functional Grammar (LFG)
  LFG - University of Essex
  LFG - Stanford University
  Grammar Writer's Workbench for Lexical Functional Grammar at Xerox PARC
- Stochastic Unification-based Grammars
  - "Stochastic Attribute-Value Grammars," Steven Abney, Computational Linguistics, number 4, volume 23, pp 597-617, 1997.

Parsing (Syntactic Analysis)
- ALE unification-based,parser - coverage: medium

- LKB unification-based,parser - coverage: medium
- PC-PATR unification-based,parser - coverage: small
- Stefy unification-based,parser - coverage: small
- NLP Software (includes parser list)
- Parser comparison (several parsers referenced)
- Collins parser, coverage: large
- Link Grammar parser, coverage: large
- Apple Pie Parser
- Probabilistic Word Graph Parser: Java Source & Documentation, Bob Carpenter, coverage: small
- MINIPAR parser, coverage: medium
- Evalb - bracket scoring program

Parse TreeBanks

- The Penn Treebank Project (English)
- NEGRA corpus (German)
- Kyoto Text Corpus (Japanese)

Machine Translation

- General
    - Machine Translation Archive - by John Hutchins
- On-line translation
    - at Dictionary.com
- MT Research
    - Building and Using Parallel Texts: Data Driven Machine Translation and Beyond HLT-NAACL 2003 Workshop, May 31, 2003
    - "Fast Decoding and Optimal Decoding for Machine Translation" (U. Germann, M. Jahr, K. Knight, D. Marcu, and K. Yamada), Proc. of the Conference of the Association for Computational Linguistics (ACL), 2001.
    - "Unification-Based Glossing," (V. Hatzivassiloglou and K. Knight), Proc. of the International Joint Conference on Artificial Intelligence (IJCAI), 1995.
- A list of Suppliers of Machine Translation Software - by The British Computer Society Natural Language Translation Specialist Group
- Systran - Comercial machine translation; available free on-line service
- Babel Fish Translation - On-line translation, Babel Fish, AltaVista (powered by SYSTRAN)

Information Retrieval

- Open Source Search Engines
    - Lucene - Jakarta Lucene is a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform. Jakarta Lucene is an open source project available for free download from Apache Jakarta. Please use the links on the left to access Lucene.
    - Wumpus - by Stefan Buettcher
- WebSPHINX - A Personal, Customizable Web Crawler
- A list of IR systems (ir.dcs.gla.ac.uk)
- System SMART
- OKAPI
- The Lemur Toolkit for Language Modeling and Information Retrieval
- Nutch search engine
- Zettair (once called Lucy)
- mg ("Managing Gigabytes")
- DataparkSearch Engine
- Lemur
- Andrew McCallum's Code and Data
- Introduction to Information Retrieval - by Chrisopher Manning, Prabhakar Raghavan, and Hinrich Schutze, 2007, draft available on-line
- Information Retrieval - A book by C. J. van Rijsbergen, 1979, available on-line
- Modern Information Retrieval - A book by Ricardo Baeza-Yates and Berthier Ribeiro-Neto, contents
- Cross-Language IR
    - CLEF - Cross Language Evaluation Forum
    - Japanese/English Cross-Language Information Retrieval: Exploration of Query Translation and Transliteration - by Atsushi Fujii and Tetsuya Ishikawa, Computers and the Humanities, Vol.35, No.4, pp.389-420, 2001.

Information Extraction

- Balie - A tool for multilingual information extraction
- Chelba, Mahajan: Information Extraction Using the Structured Language Model
- BioCreAtIvE Critical Assesment of IE system in Biology

Semantic Annotation

- Semantic Web
    - W3C Page on Semantic Web
    - W3C - World Wide Web Consortium
    - OWL - Web Ontology Language
    - DAML - DARPA Agent Markup Language

RDF - Resource Description Framework
RDF Test Cases
Mindswap - by Jim Hendler, University of Maryland
SemanticWeb.org
SWRC - Semantic Web Research Community Ontology
KSL Reports, Stanford
openRDF.org - home of Sesame
SeRQL query language
SPARQL Query Language for RDF
- Genomics
YAGI (Yet Another Gene Identifier)
LingPipe - by Alias-i
AbGene
BioNLP Resources - by Alex Morgan
- Semantic Annotation: Other
"Finding Errors Automatically in Semantically Tagged Dialogues" - by John Aberdeen, Christine Doran, Laurie Damianos, Samuel Bayer, Lynette Hirschman, The MITRE Corporation, 2001.
- Semantic Role Labeling
CoNLL-2005 Shared Task - Semantic Role Labeling
- XML-Related
oXygen XML Editor - XML Editor and XSLT Debugger
Ontologies
- SWRC - Semantic Web Research Community Ontology
- Open Cyc
- Cycorp
- SUMO - Suggested Upper Merged Ontology - link suggested by Adam Pease [Link1]
- Protege Ontologies Library
- SUMO translation for Protege frame system
- SUO
- Dublin Core
Question Answering
- TREC QA
TREC
TREC-8 Proceedings (1999)
TREC-9 Proceedings (2000)
TREC-10 Proceedings (2001)
TREC-11 Proceedings (2002)
- QA Systems
AnswerBus - Question Answering system
Start - NL Question Answering System (on-line)
HITIQA: High-Quality Interactive Question Answering, by Token Stzalkowski et al
Webclobedia by ISI, University of Southern California
- FAQ Collections
Internet FAQ Archives
Leeds University FAQ
NLP Tools
- GATE - General Architecture for Text Engineering, used in Information Extraction
- MedLEE
- OpenNLP - Open source NLP, project umbrella
- Natural Language Toolking in Pyton, nltk.sourceforge.net
- FreeLing
- The Festival Speech Synthesis System - by the Centre for Speech Technology Research, Univ.of Edinburgh
- Festival at CMU
- PRAAT: Phonetics and Speech Tools
- The EMU Speech Database System
- Commercial tools
Conexor
NL Corpora and Other NL Resources
- Standards
CES - Corpus Encoding Standard
ISO/TC 37/SC 4, Standards for Language Resources
- Word lists
List of French first names
Several word lists at project GutenMark (En, Fr, Ge, Latin, Italian, Spanish, Swedish, Finnish, ...)
- NL Corpora - Free
WAP data from the project WebACE - by Daniel Boley, see README.html before use
Project Gutenberg

- Alex Catalogue of Electronic Texts
- Russian novels
- Hansard - Parallel English/French Corpus - Official records of Canadian Parliament
    - Canadian Parliament site
    - Ulrich Germann's site
- E-mail corpora
    - Enron Email Dataset
    - Corpus of the W3C lists for TREC-Enterprise 2005 - processed at the University of Maryland
- OPUS - and open source parallel corpus
- EUconst - European Constitution in 21 languages
- Knorpora - CD for students of corpus-based computational linguistics
- Europarl Parallel Corpus - Europarl Parallel Corpus
- Archive.org texts

- NL Corpora - Free with Licence agreement
    - ICE (International Corpus of English) - University College London
- NL Corpora - Commercial
    - The LDC Corpus Catalog
    - ELDA - Evaluations and Language resources Distribution Agency

Commercial Links

- NLP Products
    - The Sketch Engine
- NLP Companies
    - Korlex
    - Lexicography MasterClass
    - TextForge
    - Language and Computing