**Presenters**

**Iryna Gurevych** (gurevych@tk.informatik.tu-darmstadt.de)
Iryna Gurevych is head of the Ubiquitous Knowledge Processing (UKP) Lab at the University of Darmstadt. Her recent research has focused on the application of lexical semantic knowledge in such areas as spoken dialogue summarization, information retrieval for educational purposes, e.g. electronic career guidance, or question answering based on question-answer repositories in Web 2.0 applied to eLearning. Her areas of expertise include algorithms for computational lexical semantics and processing of user generated discourse. She guided the development of the high-performance Java-based Wikipedia and Wiktionary APIs as well as projects in collaborative annotation, information filtering and sentiment analysis for eLearning.

**Delphine Bernhard** (delphine@tk.informatik.tu-darmstadt.de)
Delphine Bernhard is Senior Researcher in the Ubiquitous Knowledge Processing (UKP) Lab at the University of Darmstadt. She obtained her PhD in 2006 from the Université de Grenoble 1, where she worked on terminology extraction from domain specific texts and unsupervised morphological analysis. Her current work focuses on enhancing question answering systems to meet the specific needs of learners. Her further research topics include processing user generated discourse and quality assessment of social media content.

**Aljoscha Burchardt** (burchardt@tk.informatik.tu-darmstadt.de)
Aljoscha Burchardt is scientific coordinator of the Center of Research Excellence “eLearning 2.0” and Senior Researcher in the Ubiquitous Knowledge Processing Lab at the University of Darmstadt. He obtained his PhD from Saarland University in 2008, where he worked in projects related to both eLearning and applied lexical semantics. His current work focuses on the use of summarization techniques to access and present multimodal learning materials in collaborative settings.
Overview

Typical Web 2.0 tools such as wikis, blogs, and podcasts have recently entered the classroom and foster interactions between learners and tutors, within the new eLearning 2.0 paradigm. As a result, eLearning 2.0 makes large amounts of eLearning discourse available for Natural Language Processing (NLP) within the field of research that we call "Educational Natural Language Processing" (e-NLP). Research on e-NLP has existed for a long time and has focused on e.g. intelligent tutoring systems (Litman & Forbes-Riley, 2006), or essay scoring (Attali & Burstein, 2006). This field of research brings together two communities: language technology on the one side and educational computing on the other side. Several workshops on "Building Educational Applications Using NLP" and related topics have already taken place at major conferences, such as HLT-NAACL 2003, COLING 2004, ACL 2005, ACL 2008 and NAACL-HLT 2009.

NLP techniques are used in many educational applications working with textual data such as intelligent tutoring systems or computer-assisted language learning. However, these applications are particularly challenging for NLP since they require an adaptation of NLP techniques to various types of discourse, e.g. tutoring dialogues, which are different from typical task-oriented spoken dialogue systems. Moreover, educational applications place strong requirements on NLP systems, which have to be robust yet accurate. Therefore, this is an important application domain and a source of innovation for both NLP and educational computing, as shown by Feng et al. (2006), Kim et al. (2006), Malioutov & Barzilay (2006) and Csomai & Mihalcea (2007), to name just a few.

In this tutorial, we will review a variety of uses of NLP in the educational domain and point to emerging trends which call for new types of applications.
<table>
<thead>
<tr>
<th>Contents</th>
<th>Slide numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction: eLearning and NLP</td>
<td>7-12</td>
</tr>
<tr>
<td>2. Automatic generation of exercises</td>
<td>14-43</td>
</tr>
<tr>
<td>a) Computer-based testing</td>
<td>15-19</td>
</tr>
<tr>
<td>b) Multiple-choice questions</td>
<td>20-23</td>
</tr>
<tr>
<td>c) Fill-in-the-blank questions</td>
<td>24-27</td>
</tr>
<tr>
<td>d) Multiple-choice cloze questions</td>
<td>28-32</td>
</tr>
<tr>
<td>e) Matching test items</td>
<td>33-35</td>
</tr>
<tr>
<td>f) Error correction questions</td>
<td>36</td>
</tr>
<tr>
<td>g) Evaluation</td>
<td>37-43</td>
</tr>
<tr>
<td>3. Assessment of learner generated discourse</td>
<td>44-102</td>
</tr>
<tr>
<td>a) Introduction</td>
<td>44-50</td>
</tr>
<tr>
<td>b) Assessing short textual answers</td>
<td>51-64</td>
</tr>
<tr>
<td>c) Essay grading</td>
<td>65-83</td>
</tr>
<tr>
<td>d) Plagiarism</td>
<td>84-102</td>
</tr>
<tr>
<td>4. Reading and writing assistance</td>
<td>103-144</td>
</tr>
<tr>
<td>a) Text readability</td>
<td>103-109</td>
</tr>
<tr>
<td>b) Document retrieval for reading practice</td>
<td>110-113</td>
</tr>
<tr>
<td>c) Text simplification</td>
<td>114-115</td>
</tr>
<tr>
<td>d) Vocabulary assistance</td>
<td>116-125</td>
</tr>
<tr>
<td>e) Spell checking</td>
<td>128-134</td>
</tr>
<tr>
<td>f) Grammar checking</td>
<td>135-140</td>
</tr>
<tr>
<td>g) Dictionary lookup</td>
<td>141-144</td>
</tr>
<tr>
<td>5. Web 2.0 and computer supported collaborative learning</td>
<td>145-171</td>
</tr>
<tr>
<td>a) Web 2.0 &amp; eLearning 2.0</td>
<td>145-153</td>
</tr>
<tr>
<td>b) NLP for Wikis</td>
<td>154-169</td>
</tr>
<tr>
<td>c) Quality of user-generated discourse</td>
<td>170-171</td>
</tr>
<tr>
<td>7. Example e-NLP application 2: educational question answering</td>
<td>188-203</td>
</tr>
<tr>
<td>8. Conclusions</td>
<td>204-210</td>
</tr>
<tr>
<td>9. Bibliography</td>
<td>Appendix</td>
</tr>
</tbody>
</table>
UKP Lab Research Topics

- Text Mining
- Natural Language Processing / Semantics
- Semantic Information Management
- Web 2.0 / Services
- eLearning 2.0
- User-generated Discourse
- Darmstadt Knowledge Processing Repository
- WikiMining

Research Projects and eLearning

- Semantic Information Retrieval (SIR)
- Sentiment Analysis in User Generated Discourse (SentAL)
- Internet der Dienste (THESEUS)
- Semantic Question Answering for eLearning 2.0 (QA-EL)
- Self-Improving Wikis
- Wikis 2.0
- Wiki-Mining
- Darmstadt Knowledge Processing Repository
- Data export
- Project specific analysis
- Semantic analysis
- Syntactic analysis
- Morphological analysis
- Linguistic preprocessing
- Data import

Outline

- Introduction: eLearning and NLP
- Automatic generation of exercises
- Assessment of learner generated discourse
- Reading and writing assistance
- Web 2.0 and computer supported collaborative learning
- Example e-NLP application: electronic career guidance
- Example e-NLP application: educational QA
- Wrap up and questions

e-NLP

- Educational Natural Language Processing
- eLearning
- NLP
- Computer-assisted learning / instruction
- Analysis and use of language by machines
**Definition**

Field of research exploring the use of NLP techniques in educational contexts

**Web 2.0 & eLearning 2.0**

- Large text repositories with user generated discourse and user generated metadata are created
- These repositories need advanced information management and NLP to be efficiently accessed
- Using these repositories to create structured knowledge bases can improve NLP

**Feedback Loop: NLP & eLearning 2.0**

Semantic knowledge

Wikis, Blogs,...

Content creation

Intuitive access

NLP

eLearning 2.0
Outline

Introduction: eLearning and NLP
Automatic generation of exercises
Assessment of learner generated discourse
Reading and writing assistance
Web 2.0 and computer supported collaborative learning
Example e-NLP application: electronic career guidance
Example e-NLP application: educational QA
Wrap up and questions

Computer-based Testing

Definition: All forms of assessment delivered with the help of computers
Also called:
• Computer Assisted/Aided Assessment (CAA)
• Adequate question types for CAA (McKenna & Bull, 1999):
  • Multiple choice questions (MCQs)
  • True/False questions
  • Matching questions
  • Ranking questions
  • Sequencing questions
  • etc.

Question Types

Objective test items
• constrained answer, to be selected among a set of alternatives
• short answer (word or phrase) in response to a question
• objective and impartial scoring
• Examples:
  • Fill-in-the-blanks questions
  • Multiple-choice questions
  • Matching questions

Subjective test items
• original answer
• variable length
• biased scoring
• Examples:
  • Short-answer essays
  • Extended-response essays
Roles of Test Items in Learning

- **Summative assessment**
  - "Assessment of learning"
  - Measuring student achievement

- **Formative assessment**
  - "Assessment for learning"
  - Active learning: encourage learners to practice and apply newly acquired knowledge by answering test items

NLP for CAA

- **Generation of questions and exercises**
  - Writing test questions, especially objective test items, is an extremely difficult and time consuming task for teachers
  - Use of NLP to automatically generate objective test items, esp. for language learning

- **Assessment and evaluation of answers to subjective test items**
  - Use of NLP to automatically:
    - Diagnose errors in short-answer essays
    - Grade essays

Automatic Generation of Test Items

- **Source data**
  - Corpora: texts should be chosen according to
    - the learner model (level, mastered vocabulary)
    - the instructor model (target language, word category)
  - Lexical semantic resources, e.g. WordNet

- **Tools**
  - Tokeniser and sentence splitter
  - Lemmatiser
  - Conjugation and declension tools
  - POS tagger
  - Parser and chunker

Multiple-Choice Questions (MCQ)

- **Choose the correct answer among a set of possible answers**

  **Example (Mitkov et al., 2006)**
  Who was voted the best international footballer for 2004?
  (a) Henry
  (b) Beckham
  (c) Ronaldinho
  (d) Ronaldo

- **Usually 3 to 5 alternative answers**
### Distractors

- **Distractors** (also distracters) are the incorrect answers presented as a choice in a multiple-choice test.
- **Generation of "good" distractors** (McKenna & Bull, 1999; Duvall)
  - Ensure that there is only one correct response for single response MCQ
  - The key should not always occur at the same position in the list of answers
  - Distractors should be grammatically parallel with each other and approximately equal in length
  - Distractors should be plausible and attractive
  - However, distractors should not be too close to the correct answer and risk confusing students

### Automatic Generation of MCQs

1. **Selection of the key**
   - Unknown words that appear in a reading (Heilman & Eskenazi, 2007)
   - Domain-specific terms:
     - Automatically extracted (Mitkov et al., 2006)
     - Present in a thesaurus, e.g. UMLS (Karamanis et al., 2006)

2. **Generation of the stem**
   - Constrained patterns (Heilman & Eskenazi, 2007): Which set of words are most related in meaning to "reject"?
   - Transformation of source clauses to stems, using transformation and agreement rules (Mitkov et al., 2006):
     - Transitive verbs require objects → Which kind of verbs require objects?

3. **Generation of the distractors**
   - WordNet concepts which are semantically close to the key, e.g. hypernyms and co-hyponyms (Mitkov et al., 2006; Karamanis et al., 2006)
   - Stem: "Which part of speech serves as the most central element in a clause?"
     - Key: "verb", Distractors: "noun", "adjective", "preposition"
   - Thesaurus-based and distributional similarity measures (Mitkov et al., 2006)
   - Other NPs with the same head as the key, retrieved from a corpus (Mitkov et al., 2006)
     - Key: "transitive verbs", Distractors: "modal verbs", "phrasal verbs", "active verbs"

### Fill-in-the-Blank Questions (FIB)

- **Also called cloze test**
- Technique which dates from 1953 (Wilson Taylor)
- Consists of a portion of text with certain words removed
- The student is asked to "fill in the blanks"
- **Objective cloze items** = multiple-choice cloze items, i.e. students are given a list of words to use in a cloze
- **Subjective cloze items** = students can choose the words
- Challenges:
  - Phrase the question so that only one correct answer is possible
  - Spelling errors in subjective cloze items
Fill-in-the-Blank Examples

- **Blank = preposition** (Source: [http://www.purl.org/net/WERTI](http://www.purl.org/net/WERTI))
  
  SANTIAGO, May 15 (Reuters) - Chile's Chilen volcanoes groaned, rumbled and shuddered on Thursday, raising new concerns among authorities firing bolts pierced the huge clouds of hot ash. Hovering ominously, it's crater.

- **Blank = verb to be conjugated** (Source: [http://www.nonstopenglish.com/exercise.asp?exid=915](http://www.nonstopenglish.com/exercise.asp?exid=915))

  Fill in the gaps with the correct tenses: Past Simple or Present Perfect

  Example: 1. (key = have already seen)


Selection of the Blanks

- Every "n-th" (e.g. fifth or eighth) word in the text (Coniam, 1997)
- Words in specified frequency ranges, e.g. only high frequency or low frequency words (Coniam, 1997)
- Words belonging to a given grammatical category (Coniam, 1997; Aldabe et al., 2006)
- Open-class words, given their POS, and possibly targeted word sense (Liu et al., 2005; Brown et al., 2005)
- Machine learning, based on a pool of input questions used as training data (Hoshino & Nakawaga, 2005)

Objective Multiple-Choice Cloze Items

Combination of a cloze item with multiple-choice answers

(adj) strange: He thought it was ______ that her mobile was switched off.

- [ ] allegiance
- [ ] sinister
- [ ] peculiar
- [ ] grieve
- [ ] virulent

(adj) strange: He thought it was peculiar that her mobile was switched off.

- [ ] allegiance
- [ ] sinister
- [ ] peculiar
- [ ] grieve
- [ ] virulent

http://www.wordlearner.com
Generation of the Distractors

- Randomly chosen in the text from which the question was generated (Hoshino & Nakagawa, 2005)
- Same POS (Coniam, 1997)
- Similar frequency range (Coniam, 1997)
- For grammar questions, use a declension or a conjugation tool to generate different forms of the key, e.g. change case, number, person, mode, tense, etc. (Aldabe et al., 2006, Chen et al., 2006)
- Common student errors in the given context (Lee & Seneff, 2007)
- Collocations: frequent co-occurrence with either the left or the right context (Lee & Seneff, 2007)
- Open class words: semantic similarity based on distributional similarity (Smith et al., 2008) or a thesaurus (Sumita et al., 2005)

The Frequency Heuristic

(Coniam, 1997)

A University of Wolverhampton researcher, Ms. Robin Iredale, commented that a __2__ of the hiring practices of 55 companies also said there was no __3__ putting a small Asian in a __4__ of authority over taller Australians.” She said: “They said __5__ workers would not like having Asians __6__ because they work too hard.”

<table>
<thead>
<tr>
<th>Item (2)</th>
<th>Option</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>driver</td>
<td>1,716</td>
</tr>
<tr>
<td>B</td>
<td>distance</td>
<td>1,717</td>
</tr>
<tr>
<td>C</td>
<td>survey [key]</td>
<td>1,715</td>
</tr>
<tr>
<td>D</td>
<td>dream</td>
<td>1,719</td>
</tr>
<tr>
<td>E</td>
<td>tree</td>
<td>1,724</td>
</tr>
</tbody>
</table>

Table 4
Word Classes and Word Frequencies in Test Items

<table>
<thead>
<tr>
<th>Item no.</th>
<th>Word (w/o key)</th>
<th>Word class tag</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>survey</td>
<td>noun</td>
<td>1,715</td>
</tr>
<tr>
<td>3</td>
<td>point</td>
<td>noun</td>
<td>299</td>
</tr>
<tr>
<td>4</td>
<td>position</td>
<td>noun</td>
<td>632</td>
</tr>
<tr>
<td>5</td>
<td>other</td>
<td>determiner</td>
<td>80</td>
</tr>
<tr>
<td>6</td>
<td>around</td>
<td>preposition</td>
<td>201</td>
</tr>
</tbody>
</table>

Verification of the Distractors

- Basic verifications:
  - there must be enough distractors
  - there must be no duplicated distractors (Aldabe et al., 2006)
- Collocations: choose distractors that do not collocate with important words in the target sentence (Liu et al., 2005; Smith et al., 2008)
- Use of the Web: if the sentence/phrase containing the distractor is frequent on the web, then the distractor should be rejected (Sumita et al., 2005)

The child’s misery would move even the most ___ heart.
(a) torpid hits(“the most torpid heart”) = 4  
(b) invidious hits(“the most invidious heart”) = 0  
(c) stolid hits(“the most stolid heart”) = 6  
(d) obdurate hits(“the most obdurate heart”) = 1240  

Student Project in the e-NLP Course at the TU Darmstadt

- Based on “Automatic generation of cloze items for prepositions” (Lee & Seneff, 2007)
- Example:
  If you don’t have anything planned for this evening, let’s go ___ a movie.
  [a] to  [b] of  [c] on  [d] null
- Tasks:
  - INPUT: sentence + key, OUTPUT: list of three distractors
  - The three distractors must each be generated taking a different approach
    - baseline: word frequencies
    - collocations
    - “creative” method, devised by the students
- Conclusion: a motivating and interesting project for students
Matching Test Items

- Task: match items in one list with response items in another list
- Kinds of elements matched:
  - Word – synonym
  - Definition – term
  - Word – antonym
  - Hypernym – hyponym
  - Historical event – date
  - etc.
- Matching test items assess a learner's understanding of relationships

Matching Test Items for Vocabulary Assessment (Brown et al., 2005)

Wordbank:
- verbose
- infallible
- obdurate
- opaque

Choose the word from the wordbank that best completes each phrase below:
1. ___ windows of the jail
2. the Catholic Church considers the Pope ___
3. ___ and ineffective instructional methods
4. the child's misery would move even the most ___ heart

Glosses for specific word senses in WordNet

Error Detection Questions

- Aim: detect and possibly correct errors, which can be marked or not
- Example (Chen et al., 2006)
  Although maple trees are among the most colorful varieties ___ in the fall, they lose its leaves sooner than oak trees.
  (A) ___ (B) ___ (C) ___ (D) ___
- Wrong statements are produced by the distractor generator
Evaluation of Generated Questions

- Student evaluation
  - Difficulty and response time
  - Comparison with results obtained for manually generated tests (Heilman & Eskenazi, 2007)

- Instructor evaluation
  - Usability: "all distractors result in an inappropriate sentence" (Liu et al., 2005; Lee & Seneff, 2007)
  - Post-editing: count how many test items are accepted, rejected or revised by instructors during post-editing (Aldabe et al., 2006; Mitkov et al., 2006)

Pre-requisites for Student Evaluation

- External assessment
  - Evaluate the linguistic and / or factual knowledge of the students before they take the test, e.g. the Nelson-Denny Reading Test, the Raven's Matrices Test, the Lexical Knowledge Battery (Brown et al., 2005)

- Self-assessment
  - Have the students assess whether they know the target word or not (Brown et al., 2005; Heilman & Eskenazi, 2007)
    - "Do you know the word 'w'?"

Item Analysis

- Investigate the quality of the test items (Zurawski, 1998)
- Quantitative item analysis:
  - Facility / Difficulty index (p): number of test takers who answered the item correctly divided by the total number of students who answered the item
  - Discrimination index (D): "does the test item differentiate those who did well on the exam overall from those who did not?"
    - Divide the students in two groups: high-scoring and low-scoring (above and below the median)
    - Compute the item difficulty index separately for both groups: \( p_{\text{upper}} \) and \( p_{\text{lower}} \)
    - Discrimination index \( D = p_{\text{upper}} - p_{\text{lower}} \)

Example

- The child's misery would move even the most _____ heart.
  - (a) torpid chosen by 7 students
  - (b) invidious chosen by 1 student
  - (c) stolid chosen by 3 students
  - (d) obdurate chosen by 15 students

- #Students: 26
- Difficulty index: \( \frac{15}{26} = 0.58 \) → neither too difficult nor too simple (recommended score: 0.5)
- Discrimination index
  - 9 out of 12 students in the high group found the correct answer
  - 6 out of 14 students in the low group found the correct answer
  - \( D = \frac{9}{12} - \frac{6}{14} = 0.75 - 0.43 = 0.32 \)
- The test item is a quite good discriminator
Item Analysis

- **Item distractor analysis**: examine the percentage of students who select each incorrect alternative, to determine if the distractors are functioning well.

  ![Distractor Analysis Data for Upper (E) and Lower (L) Scoring Students]

<table>
<thead>
<tr>
<th>Item 1</th>
<th>Item 2</th>
<th>Item 3</th>
<th>Item 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>a, b, c, d</td>
<td>a, b, c, d</td>
<td>a, b, c, d</td>
<td>a, b, c, d</td>
</tr>
<tr>
<td>24, 321</td>
<td>1, 120</td>
<td>2, 132</td>
<td>13, 72</td>
</tr>
<tr>
<td>1, 07</td>
<td>7, 6</td>
<td>8, 1</td>
<td>11, 7</td>
</tr>
<tr>
<td>9, 51</td>
<td>5, 5</td>
<td>1, 3</td>
<td>2, 3</td>
</tr>
</tbody>
</table>

  Note: Correctly keyed alternatives for each item is identified in unbolded print.

Source: (Zurawski, 1998)

Efficiency of the Automatic Generation of Test Items

- Even though automatically generated test items have to be post-edited, this is still a lot faster than writing new test items from scratch.

  Mitkov et al. (2006) report the following figures:
  - an average of 1 minute and 40 seconds was needed to post-edit a test item in order to produce a worthy item.
  - an average of 6 minutes was needed to manually produce a test item.

Summary

- The generation of questions and exercises is actually **semi-automatic**: the system's output has to be verified and modified by an instructor.
- However, NLP-based systems considerably **reduce the time spent** by instructors to write test items, even if they have to manually correct the generated test items.
- A great variety of **NLP technologies and resources** have been successfully used so far:
  - POS tagging and parsing
  - Word sense disambiguation
  - Term extraction
  - ...

Outline

1. Introduction: eLearning and NLP
2. Automatic generation of exercises
3. Assessment of learner generated discourse
4. Reading and writing assistance
5. Web 2.0 and computer supported collaborative learning
6. Example e-NLP application: electronic career guidance
7. Example e-NLP application: educational QA
8. Wrap up and questions
Assessment of Learner Generated Discourse

- Discourse ≈ Utterance longer than a sentence
- Language form: written or spoken
- Types of learner generated discourse:
  - Emerging in institutional settings, e.g. solutions to exercises
  - Emerging in informal settings, e.g. discussions in forums (next section)

Importance of Institutional eAssessment

- Feedback to the student about her level of knowledge
- Feedback to the instructor about the progress of students’ learning
- Incentive to study certain things, to study them in certain ways, to master certain skills
- Formal means for grading and/or making a pass/fail decisions

Importance of Free-Text Assessments

- Advantages over traditional multiple-choice assessments (Bennett & Ward, 1993)
- Major obstacle is the large cost and effort required for scoring
- Automatic systems:
  - Reduce these costs
  - Facilitate extended feedback to students

Learning Exercise Spectrum Model (Bailey & Meurers 2008)

- Proposed in the context of language learning (ICALL), but applicable to different topics
Tasks Discussed in this Tutorial

- MC-Tests
- FIB
- Assessing short textual answers
- Essay grading
- Detecting plagiarism

Relating Properties of the Tasks with NLP Techniques

- Assessing short textual answers
  - Gold-standard answers can be provided
  - Specific information must be complete and correct
  - Word meaning (predicate-argument-structure) matters
  - Ressource-based apprs.

- Essay grading (Detecting plagiarism)
  - Unpredictable (no correct answer)
  - Holistic (overall organization, style, etc.)
  - Rhetorical structure matters
  - Corpus-based approaches
  - Supervised approaches

Automatic Assessment

- Diagnosis, i.e., content assessment (CAM) on learner data
  - Language learning (Bailey and Meurers, 2008)
  - Error detection in C-rater (Leacock, 2004)

- Scoring of learner data (later)
  - Essays
  - Plagiarism
  - Speech
Detecting Meaning Errors (Bailey and Meurers, 2008)

- Analysis of responses to short-answer comprehension tests
- 1-3 sentences in length
- Error codes:
  - Necessary concepts left out of learner response
  - Response with extraneous, incorrect concepts
  - An incorrect blend/substitution (correct concept missing, incorrect one present)
- Multiple incorrect concepts
- Human disagreement in 12%, eliminated from the evaluation data

**CUE:** What are the methods of propaganda mentioned in the article?

**TARGET:** The methods include use of labels, visual imagery, and beautiful or famous people promoting the idea or product. Also used in linking the product to concepts that are admired or desired and to create the impression that everyone supports the product or idea.

**LEARNER RESPONSE:**
- A number of methods of propaganda are used in the media.
- Positive or negative labels.
- Giving positive or negative labels. Using visual imagery.arioing a beautiful or famous person to promote. Creating the impression that everyone supports the product or idea.

Technology of CAM

- Input:
  - Learner’s response, one + target responses, question, source reading passage
- Linguistic analysis: annotation, alignment, diagnosis

<table>
<thead>
<tr>
<th>Annotation Task</th>
<th>Language Processing Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence Detection, Tokenization</td>
<td>MontyLiang (Liu, 2004)</td>
</tr>
<tr>
<td>Lemmatization</td>
<td>PK-KBMG (Aarts, 1993)</td>
</tr>
<tr>
<td>Spell Checking</td>
<td>Edit Distance (Levenshtein, 1966), SCOWL word lists (Arkhipov, 2004)</td>
</tr>
<tr>
<td>Part-of-Speech Tagging</td>
<td>TreeTagger (Schmid, 1994)</td>
</tr>
<tr>
<td>Noun-Phrase Checking</td>
<td>CASS (Almery, 1997)</td>
</tr>
<tr>
<td>Lexical Relations</td>
<td>WordNet (Miller, 1995)</td>
</tr>
<tr>
<td>Similarity Scores</td>
<td>PPL-IR (Tenney, 2001; Mihalcea et al., 2006)</td>
</tr>
<tr>
<td>Dependency Relations</td>
<td>Stanford Parser (Klein and Manning, 2003)</td>
</tr>
</tbody>
</table>

Source: (Bailey & Meurers, 2008)

Spell Checking Example (from Leacock & Chodorow, 2003)

- 67 different variants of Reagan in about 9,000 responses. Below are all the spelling variants of Reagan that occurred more than once:

- `Regan, Reagon, Reagen, Regans, Regean, Reagons, Ragan, Reagan, Reagen, Raegan, Ragan, Raegon, Raegan, Reagons, Raygon, Rigan, Regans, Regan, Raegon, Reagons, Regean`  

- Spell checking not as easy a task as one would think
- `Regan` is close (in terms of edit distance) to the existing word `Regan`
- Yet, in the domain of US presidents, `Reagan` is more probably the intended word

Technology of CAM

- Alignment maps new concepts from learner’s response to those in target
- Token level (abstraction from string to lemma, semantic type (e.g. date, location)
- `Houses => house => LOC`
- Chunk level, e.g., `home = his house`
- Relation level (dependency, lexical)
- Pronoun resolution

- Diagnosis analyzes if the learner’s response contains content errors

**Source:** (Bailey & Meurers, 2008)
Technology of CAM

- Given the alignment analysis, when is a learner input correct / faulty / wrong?
- Evaluation
  - Hand-written rules 81% on the development data, 63% on the test data
  - Machine learning (TiMBL), 88% accuracy on the test data for binary semantic error detection task
- Viable results

C-Rater (Leacock & Chodorow, 2002)

- Measures student understanding with little regard to writing skills
- Example question (4th grade math question used in the National Assessment for Educational Progress (NAEP)):

A radio station wanted to determine the most popular type of music among those in the listening range of the station. Would sampling opinions at a country music concert held in the listening area of the station be a good way to do this?

☐ YES  ☐ NO

Explain your answer.

Technology of C-Rater

- Content expert develops a scoring guide
- Gold standard responses
- Recognizing the equivalence of the response to the correct answers
  - Essentially paraphrase recognition
- Analysis in terms of:
  - regularizing over morphological variation
  - matching on synonyms or similar words
  - resolving the spelling of unrecognized words
  - resolving the referent of any pronouns in the response
  - predicate argument structure
- Mapping canonical representations to those of the gold standard responses
- Rule-based

Predicate Argument Structure in C-Rater

- Transform text to tuples (verbs and their arg.s): „atomic meaning units“

<table>
<thead>
<tr>
<th>Score</th>
<th>Sentence and tuple</th>
</tr>
</thead>
<tbody>
<tr>
<td>Credit</td>
<td>Most people at the country show would say that country music is the most popular music.</td>
</tr>
<tr>
<td></td>
<td>say, subject most people, subject most popular music</td>
</tr>
<tr>
<td>Credit</td>
<td>The people at the country concert would all answer country music.</td>
</tr>
<tr>
<td></td>
<td>answer, subject people, subject country music</td>
</tr>
<tr>
<td>Credit</td>
<td>People at a country concert might think that country music is the best music.</td>
</tr>
<tr>
<td></td>
<td>think, subject people, subject best music</td>
</tr>
<tr>
<td>No credit</td>
<td>I happen to like country music and so do most of my friends.</td>
</tr>
<tr>
<td></td>
<td>like, subject I, subject most of my friends, subject country music</td>
</tr>
</tbody>
</table>
Problems with this Simple Approach to Predicate Argument Structure (Excursion)

- Variation in language is much more pervasive
- Simple example: passive voice
  - Mary ate the cake. (subject: Mary)
  - The cake has been eaten by Mary. (subject: the cake)
- Simple solution: check for passive (syntactic parser) and switch arguments

- Harder example:
  - John is afraid of Ghosts.
  - Ghosts scare John.
- Solution: Use a semantic resource like FrameNet.

Frame Semantics and FrameNet
(Fillmore 1976, Baker et. al. 1998)

- Lexical semantic classification of predicates and their argument structure
- A frame represents a prototypical situation (e.g. Commercial_transaction, Theft, Awareness)
- A set of roles identifies the participants or propositions involved
- Frames are organized in a hierarchy
- Berkeley FrameNet Project db: 600 frames, 9,000 lexical units, 135,000 annotated sentences

Linguistic Normalization
(Frame: Commerce_buy)

<table>
<thead>
<tr>
<th>Role</th>
<th>Example Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seller</td>
<td>BMW bought Rover from British Aerospace.</td>
</tr>
<tr>
<td>Buyer</td>
<td>Rover was bought by BMW, which acquired [...] the new Range Rover</td>
</tr>
<tr>
<td>Goods</td>
<td>BMW, which acquired Rover in 1994, is now dismantling the company.</td>
</tr>
<tr>
<td>Money</td>
<td>BMW's purchase of Rover in 1994 was a good move.</td>
</tr>
</tbody>
</table>

Voice: active / passive
POS: verb / noun

Wrapping up Content Analysis

- Applicable for short, predictable answers
- Usually resource-based
  - Spell-checkers, Grammars
  - Semantic resources
  - Special rule-based systems
  - ...
- A Result of a c-rater experiment (Leacock and Chodorow 2003)
  - About 84% agreement with human judgment
  - 47% baseline for majority class (full / partial / no credit)
Tasks Discussed in this Tutorial

- MC-Tests
- FIB
- Assessing short textual answers
- Essay grading
- Detecting plagiarism

What is an Essay?

- A major part of formal education (at least in the USA)
- Secondary students are taught structured essay formats to improve their writing skills
- Often used by universities in selecting applicants
  - Students are asked to explain, comment on, or assess a topic of study
  - These admission essays are used to judge the mastery and comprehension of the material

Essay Prompts

- Descriptive prompt:
  - “Imagine that you have a pen pal from another country. Write a descriptive essay explaining how your school looks and sounds, and how your school makes you feel.”

- Persuasive prompt:
  - “Some people think the school year should be lengthened at the expense of vacations. What is your opinion? Give specific reasons to support your opinion.”


Research Development in Writing Evaluation

Most Prominent Systems

- **Intelligent Essay Assessor** (Landauer, Foltz & Laham, 1998)
  - Based on a statistical technique for summarizing the relations between words in a document, i.e. every word is a “mini-feature”
- **Intellimetric** (Elliot, 2001)
  - Based on hundreds of undisclosed features
- **Project Essay Grade – PEG** (Page, 1994)
  - Based on dozens of mostly undisclosed features
- **E-Rater** (Burstein et al., 1998)
  - The 1st version used more than 60 features
  - E-rater 2.0 uses a small set of features

How Do Humans and Machines Rate Essays?

- Humans evaluate various **intrinsic variables** of interest → essay score:
  - Content adequacy
  - Structure
  - Argumentation
  - Diction
  - Fluency
  - Correct language use

- Machines use **approximations** or **possible correlates** of intrinsic variables → scoring model

How is a Scoring Model Created?

- Analyze a few hundred essays:
  - Written on a specific prompt
  - Pre-scored by as many human raters as possible

- Identify most useful approximations (classification features) out of those available to the system

- Employ a statistical modeling procedure to combine the features and produce a machine-generated score

Validating the Meaning of Scores (Yang et al. 2002)

- Relationship between human and machine scores of the same prompt:
  - Compare the machine-human and human-human agreement (Burstein et al., 1998; Elliot, 2001; Landauer et al., 2001)
  - Estimate a true score as the one assigned by multiple raters (Page, 1966)

- Relationship between test scores and other similar measures:
  - Compare automatic scores with multiple-choice test results and teacher judgments (Powers et al., 2002)

- Understanding the scoring process, i.e. relative importance of different writing dimensions:
  - Most commonly used features in scoring models (Burstein et al., 1998)
  - The most important component is content (Landauer et al., 2001)
Skepticism and Criticism
(Page and Petersen, 1995)

- Three general directions of criticism:
  - **Humanistic** – never understand or appreciate an essay as a human
    → Use automatic scoring as a second rater
  - **Defensive** – playful or hostile students produce “bad faith” essays
    → a study by Powers et al. (2001), a lot of data needed
  - **Constructive** – computer-measured variables is not what is really important for an essay
    → an improved ability to additionally provide diagnostic feedback

Features Used by e-Rater 2.0
(Burstein et al., 1998)

- Measures of:
  - Grammar, usage, typos
  - Style
  - Organization & development
  - Lexical complexity
  - Prompt-specific vocabulary usage

- Implemented in different writing analysis tools

- Based on an NLP foundation that provides instructional feedback to students in the web-based Criterion system

Writing Analysis Tools: Correctness

- Identify five main types of grammar, usage and mechanics errors:
  - Agreement and verb formation errors, wrong word use, missing punctuation, typographical errors

- Corpus-based approach:
  - Train the system on a large corpus of edited text
  - Extract and count bigrams of words and POS
  - Search for bigrams in essay that occur much less often (Chodorow & Leacock, 2000)
  - *girl walk* occurs less frequently than *girl walks*

Writing Analysis Tools: Aspects of Style

- The writer may wish to revise:
  - The use of passive sentences
  - Very long or very short sentences
  - Overly repetitious words (Burstein & Wolska, 2003)
Writing Analysis Tools:
Organization & Development

- Discourse elements present or absent in the essay (Burstein, Marcu and Knight, 2003)
- A linear representation of text as a sequence of:
  - Introductory material
  - A thesis statement
  - Main ideas
  - Supporting ideas
  - A conclusion
- How can we find these parts automatically?

Supervised Learning

- Train a system on a large corpus of human annotated essays to identify "good" sequences
- The computer extracts regularities such as:
  - Mandatory parts,
  - Number restriction, e.g., > 3 main ideas,
  - …

Essay Annotated with Discourse Elements


Writing Analysis Tools:
Lexical Complexity

- Related to word-specific characteristics such as:
- A measure of vocabulary-level, based on Breland, Jones and Jenkins (1994), *Standardized Frequency Index* across the words in an essay
- The average word length in characters in an essay
Writing Analysis Tools: Prompt-Specific Vocabulary Usage

- Intuition: good essays resemble each other in their word choice, as will poor essays (within the same prompt)
- Idea: compare an essay to a sample of essays from each score category (usually 1-6)
  - Each essay and a set of training essays from each score category is converted to a vector
  - Some function words are removed
  - Each vector element is a weight based on a word frequency function
  - Six cosine correlations are computed between the essay and each score category to determine the similarity

Scoring in e-Rater 2.0

- Input: all features of all writing analysis tools
  - Grammar, usage, mechanics, style (4 features)
  - Organization & development (2 features)
  - Lexical complexity (2 features)
  - Prompt-specific vocabulary usage (2 features)
- Straightforward combination method:
  - Apply a linear transformation on feature values to achieve a desired scale
  - A weighted average of the standardized feature values

Future Directions

- Better standardization of scoring - a single scoring model for all prompts of a program or assessment
- Better understanding and control over the automated scores
- Cover more aspects of writing quality, devise new features
- Prefer features providing useful instructional feedback
- Detection of anomalous and bad-faith essays
- Characterize different types of anomalies
- Detect off-topic essays (Higgins, Burstein and Attali, 2006)

Tasks Discussed in this Tutorial

- MC-Tests
- FIB
- Assessing short textual answers
- Essay grading
- Detecting plagiarism
Plagiarism

“Plagiarize: [...] to take and use as one’s own the thoughts, writings, or inventions of another. [...]”

Oxford English Dictionary Online

• Main Feature: Missing indication of source

Affected Types of Media

• Music
• Text
• Graphics
• Images
• ...

In this context: written text

Plagiarism at Universities

• Intra-corpal plagiarism
  • Copying from fellow students
  • Kollusion (here: unwanted group work)

• Web-based plagiarism
  • Copying from an online source (book, web page, etc.)

(Culwin and Lancaster 2001)

• „Web 2.0-mentality“: Find-Remix-Share
  
(Sattler 2007)

Plagiarism at Universities (Lecturers/Researchers)

• In teaching material: slides / course reader / etc.

• Self-plagiarism
• Silent inclusion of results in one’s own work (from PhD candidates, students, etc.)

(http://www.spiegel.de/unispiegel/jobundberuf/0,1518,207062,00.html)

• Peer-Reviews (project proposals, conference papers)

(http://de.wikipedia.org/wiki/Plagiat#Plagiate_in_Hochschule_und_Schule)

• Honorary authorship
Types of Plagiarism

1. Plagiarism of authorship: the direct case of putting your own name to someone else’s work.
2. Word-for-word plagiarism: copying of phrases or passages from a published text without quotation or acknowledgement.
3. Paraphrasing plagiarism: words or syntax are changed (rewritten), but the source text can still be recognized.
4. Plagiarism of the form of a source: the structure of an argument in a source is copied (verbatim or rewritten).
5. Plagiarism of ideas: the reuse of an original thought from a source text without dependence on the words or form of the source.
6. Plagiarism of secondary sources: original sources are referenced or quoted, but obtained from a secondary source text without looking up the original.

Based on Martin (1994) and Clough (2003)

Typical Plagiarism Indicators

- Use of advanced or technical vocabulary beyond that expected of the writer.
- A large improvement in writing style compared to previous submitted work.
- Inconsistencies within the written text itself, e.g. changes in vocabulary, style (e.g. references) or quality.
- Incoherent text where the flow is not consistent or smooth.
- Dangling references: a reference appears in the text, but not in the bibliography and vice versa.
- A large degree of similarity between the content, mistakes, etc. of two or more submitted texts.

Based on Clough (2003)

Techniques Used to Conceal Copying

- Replacing odd or unusual words.
- Changing formatting.
- Adding filler words or phrases.
- Changing headings.
- Rephrasing sentences.
- Removing or re-ordering sections.
- Changing spelling (usually from American English to British English, if the document is plagiarized from the Web).
- Producing consistency by find-and-replace (as an example, if some papers refer to the World Wide Web, some to the WWW, some to the Web, a student may perform a global find-and-replace to ensure consistency within the plagiarized document).
- In programming, changing variable names and comments.

The use of electronic tools to support plagiarism detection:
http://www.comp.leeds.ac.uk/hannah/CandIT/plagiarism.html

String Matching Algorithms

- Most popular plagiarism detection scheme:
  - Comparing word windows of length ≥ n.
  - Computing the overlap of matching subsequences and substrings (consecutive tokens).
  - n is derived empirically.
  - The longer n becomes, the more unlikely it is that the same sequence will appear in independently written texts.
  - Problem: larger n-grams types are rare, difficult to define thresholds.
Uniqueness of N-grams
(from Clough 2003)

- Figures taken from 769 texts in the METER corpus:

<table>
<thead>
<tr>
<th>N (words)</th>
<th>N-gram occurrences (tokens)</th>
<th>Distinct n-grams (types)</th>
<th>% distinct n-grams</th>
<th>% distinct n-grams in 1 tile</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>137204</td>
<td>14417</td>
<td>12</td>
<td>39</td>
</tr>
<tr>
<td>2</td>
<td>248819</td>
<td>96952</td>
<td>46</td>
<td>67</td>
</tr>
<tr>
<td>3</td>
<td>248819</td>
<td>180674</td>
<td>73</td>
<td>82</td>
</tr>
<tr>
<td>4</td>
<td>257312</td>
<td>214119</td>
<td>85</td>
<td>50</td>
</tr>
<tr>
<td>5</td>
<td>251429</td>
<td>225369</td>
<td>90</td>
<td>53</td>
</tr>
<tr>
<td>6</td>
<td>250956</td>
<td>231800</td>
<td>92</td>
<td>64</td>
</tr>
<tr>
<td>7</td>
<td>250306</td>
<td>234600</td>
<td>94</td>
<td>55</td>
</tr>
<tr>
<td>8</td>
<td>249954</td>
<td>236310</td>
<td>95</td>
<td>66</td>
</tr>
<tr>
<td>9</td>
<td>248841</td>
<td>237409</td>
<td>96</td>
<td>67</td>
</tr>
<tr>
<td>10</td>
<td>289610</td>
<td>278503</td>
<td>95</td>
<td>67</td>
</tr>
</tbody>
</table>

Table 1: Uniqueness of consecutive n-word sequences (n-grams) as n increases from 1-10 words

Longest Common Substrings
Computed between Two Sentences

- Greedy String Tiling (or GST; see, e.g. (Wise, 1993)), an algorithm which computes a maximal mapping of text pairs with non-overlapping substrings (called tiles).
- Advantage: n-gram size needs not be set a priori

The output of the GST algorithm is a list like: [for two years], [driver who], [into the], [a], [queen], [was] and [banned].

- Different quantitative measures can then be applied, e.g.:
  - the minimum and maximum tile length
  - the average tile length
  - the dispersion of tile lengths

Goal: derive a similarity measure for plagiarism

Challenge: distinguish derived and non-derived text(s)

Example of Tiling for Derived and Non-Derived Text (from Clough 2003)

- It has been empirically found that:
  - derived texts (top) share longer matching substrings
  - the tiling for a derived and non-derived text pair are in most cases apparently different
Machine Learning in Plagiarism Detection

- Input: Documents and their features (Document length, match size, etc.)
- Goal: A computational model that distinguishes original and plagiarism

- **Supervised (machine) learning**: train a classifier on manually annotated training data (texts classified as plagiarized or not)
  - Disadvantage: Many documents needed (thousands)
- **Unsupervised learning**: have the machine find certain “clusters”
  - Concrete instruction: Divide these texts in two parts (given these features)
  - Hope: one part will contain originals and one part derived texts
- Evaluation: check random samples

Relaxing the Approach

Preserving longer matching n-grams and tile lengths to make the approach resistant to simple edits

- Allow small gaps to represent token deletion
- Allow simple word substitution (using WordNet)
- Allow insertion of certain words such as domain-specific terminology and function words (e.g. conjunctions)
- Allow simple reordering of tokens (e.g. transposition)

NLP in Plagiarism Detection

- Existing work involves minimal natural language processing (NLP)
- Areas of NLP that could aid plagiarism detection, particularly in identifying texts which exhibit similarity in semantics, structure or discourse, but differ in lexical overlap and syntax
- NLP methods include:
  - morphological analysis, part-of-speech tagging, anaphora resolution, parsing (syntactic and semantic), co-reference resolution, word sense disambiguation, and discourse processing
- Future work:
  - several similarity scores based on lexical overlap, syntax, semantics, discourse and other structural features

How to Avoid Plagiarism?

- Clearly define plagiarism to the students and use explicit examples
- Educate the students about the honor code and the ramifications if it is violated
- Create assignments that make plagiarism difficult
- Make sure the students are familiar with online resources
- Have the students submit evidence of the research process as well as the paper
- Avoid repeating assignments and paper topics
- Inform the students you are Internet savvy and you know about the paper mills (visit the sites with the students to evaluate the quality of the work)
- Inform the students that you use plagiarism detection software

Online Internet Plagiarism Services

- Plagiarism.org [www.plagiarism.org](http://www.plagiarism.org)
  - The largest online plagiarism service available
- None of the services details their implementation details
- All of them are commercial, but plagiarism.org allows free trial

Summing up

- Resource-based vs. corpus-based approaches
- Resources: spell checker, grammar, thesaurus, semantic net, ...
- Corpus-based approaches
  - Supervised: Manual annotation and generalization
  - Unsupervised: Automatic induction of structure

Outline

Introduction: eLearning and NLP
- Automatic generation of exercises
- Assessment of learner generated discourse
- Reading and writing assistance
- Web 2.0 and computer supported collaborative learning
- Example e-NLP application: electronic career guidance
- Example e-NLP application: educational QA

Readability

- "Readability is what makes some texts easier to read than others" (DuBay, 2004)
- Heavily dependent on the intended audience
- A text's readability can be estimated with readability formulas, which provide an objective prediction of text difficulty, usually expressed in terms of school grade level
- Aims:
  - match reading materials with the abilities of the readers
  - support authors in writing clearly understandable texts
### Traditional Readability Measures

<table>
<thead>
<tr>
<th>Formula</th>
<th>Date</th>
<th>Features</th>
<th>Example values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flesch index</td>
<td>1948</td>
<td>- average # syllables / word</td>
<td>- 30 = &quot;very difficult&quot;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- average sentence length</td>
<td>- 70 = &quot;easy&quot;</td>
</tr>
<tr>
<td>Fog index</td>
<td>1952</td>
<td>- # words with more than 2 syllables</td>
<td>- 6 = comic books</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- average sentence length</td>
<td>- 10 = newspapers</td>
</tr>
<tr>
<td>SMOG grading</td>
<td>1969</td>
<td>- # words with more than 3 syllables</td>
<td>- 0 to 6 = low-literate</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- 19+ = post-graduate</td>
</tr>
</tbody>
</table>

### Readability Statistics

- Computed using the `style` command

- **Rotkäppchen**
  - DIEZEIT

### Statistical Language Models for Reading Difficulty

- Use of statistical models representing norms, specific populations and individuals (Brown & Eskenazi, 2004)
- Different models can be created for each level of reading difficulty (Collins-Thompson & Callan, 2005)
- Method (Collins-Thompson & Callan, 2005; Heilman et al., 2007, 2008):
  - For a given text passage $T$, the semantic difficulty of $T$ relative to a specific grade level $G_i$ is predicted by calculating the likelihood that the words of $T$ were generated from a representative language model of $G_i$
  - Reading difficulty = grade level of the language model most likely to have generated the passage $T$

### Readability analysis as a classification task

- Aim: label texts with grade levels
- Method: train multiple classifiers on manually annotated text
  - Linear regression (Feng et al., 2009)
  - Support vector machines (Petersen & Ostendorf, 2009)
- Features:
  - Lexical features: avg. number of words per sentence, avg. number of syllables per word
  - Syntactic features: parse tree height, noun phrase count, verb phrase count, SBAR count
Discourse features

- Discourse features (Pitler & Nenkova, 2008):
  - Vocabulary and discourse relations are the strongest predictors of readability (Wall Street Journal texts)
  - Discourse relations also robustly predict readability rankings (comparisons between two documents)

- Cognitively motivated features for a specific group of users (Feng et al., 2009)
  - Target group: adults with intellectual disabilities
  - Discourse level features: entity density, lexical chains

Document Retrieval for Reading Practice

- Reading proficiency is a widespread problem
  - 29% of high school seniors in public schools across America were below basic achievement in reading in 2005 (Miltsakaki & Troutt, 2008)
  - Low reading proficiency may have dramatic consequences (DuBay, 2004):
    - The strongest risk factor for injury in a traffic accident is the improper use of child safety seats
    - 79 to 94% of car seats are used improperly
    - Installation instructions are too difficult to read for 80% adult readers in the US

- Use readability measures to identify suitable and authentic documents, given a reader profile / reading grade

Vygotsky's Zone of Proximal Development

- Materials for assisted reading should be harder than the reader's tested reading level, but within the zone of proximal development

- Materials for unassisted reading, e.g. medicine inserts, instructions, should be as easy as possible

Read-X (Miltsakaki & Troutt, 2008)

- http://net-read.blogspot.com/
REAP search (Heilman et al., 2008)

Text Simplification

- The readability of a text can be improved by transforming it into a simpler text.
- Characteristics of manually simplified texts (Petersen & Ostendorf, 2007):
  - shorter sentences
  - fewer and shorter phrases
  - fewer adjectives, adverbs and coordinating conjunctions
  - nouns are less often replaced with pronouns

Original text: Congress gave Yosemite the money to repair damage from the 1997 flood.
Abridged text: Congress gave the money after the 1997 flood

Automatic Text Simplification

- Related techniques: summarisation and sentence compression
- Syntactic simplification:
  - Removal or replacement of difficult syntactic structures, using hand-built transformational rules applied to dependency and parse trees (Carroll et al., 1999; Inui et al., 2003)
- Lexical simplification:
  - Goal: replace difficult words with simpler ones (Carroll et al., 1999; Lal & Rüger, 2002)
  - Difficult words are identified using the number of syllables and/or frequency counts in a corpus
  - Choose the simplest synonym for difficult words in WordNet

Vocabulary Assistance for Reading

- Overall goal: support vocabulary acquisition during reading for:
  - children, who learn to read (Aist, 2001)
  - foreign language learners, who read texts in a foreign language
- Problem: a word’s context may not provide enough information about its meaning
- Solution: augment documents with dynamically generated annotations about (problematic) words
Selection of Target Words

- All words are annotated
- Annotate selected words
  - Manually selected target words
  - Automatically selected target words
- (Aist, 2001):
  - Words with few senses in WordNet (to avoid WSD)
  - Not a trivially easy word: three or more letters long, not in a stop list of function words, not a number
  - Not a proper noun
  - Socially acceptable, e.g. no secondary slang meanings
- (Mihalcea & Csomai, 2007): keyword extraction methods

Resources for Vocabulary Assistance

- WordNet (Aist, 2001):
  - Extraction of comparison words for a target word: antonym, hypernym, synonym
- Generation of factoids:
  - eggshell can be a kind of natural covering
- Problems:
  - some of the automatically generated factoids are too obscure or do not match the sense of the word used in the original text
  - some of the comparison words may be harder to understand than the target word
  - hypernyms do not always capture the key elements of the meaning of a word

Resources for vocabulary assistance

- Collaborative and online resources, e.g. Wikipedia, Wiktionary, Beolingus, ...

Wikipedia and Wiktionary as Lexical-Semantic Resources

http://lingro.com/
Wiktionary as Lexical-Semantic Resource

- Language
- Etymology
- Pronunciation
- Part-of-speech
- Word senses
- Synonyms
- Derived terms
- Translations

- Abbreviations, Antonyms, Categories, Collocations, Examples, Glosses, Hypernyms, Hyponyms, Morphology, Quotations, Related terms, Troponyms

JWKTL – Wiktionary API

- Freely available for research purposes
- http://www.ukp.tu-darmstadt.de/software/

Wikify! (Mihalcea & Csomai, 2007)

- Aim: link keywords (important concepts) in a document to the corresponding Wikipedia page
- Keyword extraction
  - Ranking: tf.idf, $\chi^2$ independence test, keyphraseness
- Word Sense Disambiguation to identify the target Wikipedia page:
  - Lesk algorithm: measure of contextual overlap between the Wikipedia page of the ambiguous word / phrase and the context where the ambiguous word / phrase occurs
  - Machine Learning classifier

Spelling Error Detection and Correction

- Aim: identify and correct spelling errors
- Types of spelling errors:
  - Non-word spelling errors
    - occurred instead of occurred
    - alter instead of after, later, alter, water, ate
  - Word conflation or splitting
    - ofthe, understandhme
    - sp ent, th ebook
  - Malapropisms: real-word spelling errors in open-class words
    - diary – dairy
    - there – their – they're
Research Problems (Kukich, 1992)

- **Non-word error detection**
  - From the early 1970s to the early 1980s
  - Focus on efficient pattern-matching and string comparison techniques
- **Isolated-word error correction**
  - Started in the early 1960s
- **Context-dependent word correction**
  - Started in the early 1980s
  - Use of statistical language models

**Textbook overviews:** (Jurafsky & Martin, 2008; Manning, Raghavan and Schütze, 2008)

---

### Non-word Error Detection

- **n-gram analysis**:
  - n-gram = n-letter sub-sequences of words or strings
  - examine each letter n-gram in an input string
  - find the n-gram in a table of n-gram statistics compiled from a corpus of text
  - highly infrequent n-grams indicate probable misspellings
  - especially useful for optical character recognition devices
- **Dictionary lookup**:
  - check if an input string appears in a dictionary of acceptable words
  - techniques: hash tables, tries, finite-state automata, Aho-Corasick algorithm, ternary search trees

---

### Isolated Word Error Correction

1) Detection of errors in single words, out of context
2) Generation of candidate corrections
   - Distance/Proximity metric between the correct word and the erroneous word
   - Minimum edit distance: minimum number of editing operations (i.e., insertions, deletions, and substitutions) needed to transform one string into another

```
levenshtein levenshtein
```

```
Distance = 4
```

- *"* Match; *o* Substitution; *+* Insertion; *- Deletion

(c) [www.levenshtein.net](http://www.levenshtein.net)

3) Ranking of candidate corrections based on the distance/proximity metric or occurrence counts

---

### Isolated Word Error Correction

**Problem:** even humans do not achieve 100% accuracy levels, given isolated misspelled strings (Kukich, 1992):

- vver → over, ever, very?
- wekk → week, well, weak?
Context-dependent Error Correction

- Also called context-sensitive spelling correction
- Aim: correct real-word spelling errors, which cannot be identified by dictionary lookup
- Between 25% and 40% of spelling errors are valid English words (Kukich, 1992)
- Use the context to help detect and correct spelling errors
- Based on language models

Spelling Correction for Foreign Language Learners (Heift & Rimrott, 2007)

- 80% of the misspellings produced by non-native writers of German are due to insufficient command of the foreign language:
  - Metz for Fleisch (from Metzger)
  - tanzed for tanzte (from danced)
- These errors are difficult to correct for generic spell checkers → need for rules that are geared towards common L2 errors
- Importance of feedback: learners are more likely to correct a mistake if the feedback contains explicit information on the error and correction suggestions

Grammar Checking

- Tasks:
  - Grammatical error detection: identify sentences which are grammatically ill-formed
  - Grammatical error correction: correct grammatically ill-formed sentences
- Methods:
  - Rule-based checking: use of manually written rules
  - Syntax-based checking: use the output of a parser
  - Statistics-based: use statistical information about n-gram frequencies
  - Many methods focus on a specific part-of-speech, e.g. prepositions

Grammatical Error Types

- According to (Nicholls, 1999, quoted by Chodorow & Leacock, 2000):
  - Insertion of an unnecessary word: *affect to their emotions
  - Deletion of a word: *opportunity of job
  - Word or phrase that needs replacing: *every jobs
  - Word use in the wrong form: *knowledges
- Grammatical difficulties for ESL learners:
  - Prepositions: *arrive to the town, *most of people, *He is fond this book (Chodorow et al., 2007)
  - Verb forms: I can't *skiing well, I don't want *have a baby (Lee & Seneff, 2008)
  - Articles
Rule-based Grammar Checking

- Analyse errors in a corpus and write rules to identify and correct these errors, based on POS information
- Rule patterns should not occur in correct sentences
- Examples:
  - Language Tool (Naber, 2003)
  - Open Source language checker
  - Rules are defined in XML configuration files and include feedback messages
  - GRANSKA (Eeg-Olofsson & Knutsson, 2003)
  - Rules expressed in a specific rule language
  - Recall = 25%, Precision = 100%

Syntax-based Grammar Checking

- Template-matching on parse trees (Lee & Seneff, 2008)
  - Automatic introduction of verb form errors in a corpus
  - Parsing of the corpus
  - Identification of templates in the "disturbed" parse trees

Statistics-based Grammar Checking

- Detection of unfrequent sequences of words and/or POS tags:
  - POS bigrams (Atwell, 1987)
  - POS tags and function words n-grams (Chodorow & Leacock, 2000)

  Machine learning:
  - Maximum entropy model trained with contextual features and combined with rule-based filters (Chodorow et al., 2007)
  - Machine learning model based on automatically labelled sequential patterns (Sun et al., 2007)

Classification based approach

- Method: train a classifier on grammatically correct text to predict which preposition / determiner is correct in a given context (Gamon et al., 2008; De Felice & Pulman, 2008)
- Example contextual features (De Felice & Pulman, 2008):

  Table 1: Determiner feature set for Pick the juiciest apple on the tree.
  Table 2: Preposition feature set for John drove to London.
The Tip of the Tongue Problem

Writers may want to look for words that express a given concept and are appropriate in a given context.

Problem: in order to access words in a traditional dictionary, you have to know the word you are looking for.

Dictionary Lookup (Ferret & Zock, 2006)

- Tip of the tongue problem:
  - domesticated animal, producing milk suitable for making cheese
  - NOT (cow, buffalo, sheep)
  - → goat
- The mental lexicon is a huge network of interconnected words and concepts
- The network is entered through the first word that comes to mind and the target word is retrieved thanks to connecting links.

Internal Representation

Wikipedia Graph
Outline

Introduction: eLearning and NLP
Automatic generation of exercises
Assessment of learner generated discourse
Reading and writing assistance
Web 2.0 and computer supported collaborative learning
Example e-NLP application: electronic career guidance
Example e-NLP application: educational QA
Wrap up and questions

Characteristics of Web 2.0

- Collective intelligence
- Huge amount of data
- Fast growing
- Noise
- Duplicates
- Content of different quality

New Learning Paradigms in eLearning 2.0

- Study at any place, any time
  - Several devices may be used for learning: computer, iPod, PDA, etc.
- Authority in educational systems is distributed: collective intelligence and wisdom of the crowds
  - Learn not only from teachers and instructors, but also from peers
- New forms of knowledge organization: tags and folksonomies

(e)Learning 2.0

- Main characteristics:
  - Worldwide learning community
  - Educational material produced both by students and teachers
- Tools:
  - Wikis
  - Blogs
  - Podcasts
  - Widgets
  - ...
"CALL 2.0"

Widgets for CALL

User contributed contents

User contributed contents
Use of Web 2.0 Resources

---

**Wikis**

- **Goal:** build and share knowledge

- Wikis allow users to change contents:
  - collaborative authoring
  - simple wiki markup language
  - stored edit history

- **Uses in education:**
  - Distribute educational material to students
  - Support student group work
  - Support teacher collaboration

---

**Wiki examples**

---

**Educational wiki**

---
Problems with Wikis

In the beginning ...

- Small
- Well structured
- Easy to find and add content

People like it and add lots of content

I can't find anything!

Where do I put this?

Disorientation and cognitive overload

Wiki User Survey at UKP

- 15 participants
- The two biggest problems
  - Wiki capabilities to re-organize content
  - Finding information
- Confirmed by other studies, e.g.

UKP's Approach: Wikulu

Use Natural Language Processing to support the user by providing suggestions while:
adding, organizing and finding content.

"Wikulu" - Hawaiian for organize ['kukulu'] fast ['wiki']

Adding Content: Detect Duplicate Content
Adding Content: Suggest Points of Insertion

- Text similarity (Gabrilovich & Markovitch, 2007)
- Highly similar documents might be duplicates
- ... or possible places for adding the new content
- Text segmentation (Choi et al., 2001)
- Find specific position for inserting new text by segmenting pages into coherent topics

Organizing Content: Suggest Tags

CHICAGO, Oct 29 - Kraft Foods Inc and Kellogg Co posted better-than-expected third-quarter profits on Wednesday as price increases and new products helped lift sales in a weak economy. Kraft also stood by its forecasts for 2008 earnings before one-time items as well as for 2009 net income, while Kellogg said its profit this year should hit the high end of its previous targeted range. Both Kraft, the largest North American food maker, and Kellogg, the world’s largest cereal company, have taken steps to cut costs and put more money into advertising. Both have also bolstered new product development to attract consumers even as rising commodity costs pushed them to raise prices. Commodities like wheat and energy have become less expensive in recent months, but food companies may not see a big benefit until next year, in part because they lock in their costs months ahead. Kraft, which makes Oreo cookies, Tang breakfast drink and Oscar Mayer hot dogs, reported a profit of 45 cents a share before one-time items, a penny above what analysts polled by Reuters Estimates had expected.
Organizing Content: 
Suggest Page Split/Merge

- Link detection (Green, 1998)
- Suggest similar content as link target
- Keyphrase extraction (Mihalcea & Tarau, 2004)
- Propose important keyphrases as possible tags
- Text segmentation
- Find coherent topics in a page to propose splits
- Text similarity
- Find scattered pages similar enough to merge

Finding Content: 
Recall-Oriented Search

Wiki
"issue"
"problem"

Finding Content: 
Show Related Pages While Browsing
Finding Content: NLP Algorithms

- Text similarity
- Improve search recall by taking into account term similarity to find additional relevant pages
- Show related pages while browsing

What is actually the Quality of Web 2.0 Resources?

- Wikipedia:
  - Open edit policy, yet high quality articles (Giles, 2005)
  - 42 entries tested by experts
  - Average science entry in Wikipedia contained around four inaccuracies
  - Average science entry in Encyclopaedia Britannica contained around three inaccuracies
- Automatic assessment of the quality of these resources:
  - Social Q&A sites (Jeon et al., 2006; Agichtein et al., 2008)
  - Wikipedia (Druck et al., 2008)
  - Forums (Weimer et al., 2007; Weimer & Gurevych, 2007)

Quality Assessment of User Generated Discourse

- Web 2.0 leads to massive amounts of data
- Users need content of good quality
- Current approach
  - Users label the data for quality
  - Labels are used for filtering
- Problems:
  - Happens rarely
  - New item problem
  - Premature negative consensus (Lampe and Resnick, 2004)

Outline

- Introduction: eLearning and NLP
- Automatic generation of exercises
- Assessment of learner generated discourse
- Reading and writing assistance
- Web 2.0 and computer supported collaborative learning
- Example e-NLP application: electronic career guidance
- Example e-NLP application: educational QA
- Wrap up and questions
The SIR project:
Semantic Information Retrieval for Electronic Career Guidance

Deutsche Forschungsgemeinschaft funded by the German Research Foundation

Electronic Career Guidance

Problem of Standard Information Retrieval

- Standard search engines
  1. Return many irrelevant documents (low precision)
  2. Miss many relevant documents (low recall)

- Why is this the case?
  - Pure keyword search is often out of context (e.g., apple, jaguar)
  - Vocabulary gap:
    - Words are confused with their meaning (car = automobile)
    - Related words are not considered

Vocabulary Mismatch Problem

Essay about professional interests

Descriptions of professions

Documents

Information Retrieval

Ranked List of Professions

Semantic Relatedness

I like baking cakes...

...pastries...
...confectioner...
...food processing industry

Profession 1

Profession 2

Profession 3
Where Does the Information Come From?

Knowledge Sources

- GermaNet: German lexical-semantic wordnet
- Nouns, verbs, adjectives
- 27,824 noun synsets, 8,810 verb synsets, 5,141 adjective synsets
- 60,646 words in synsets
- Wikipedia
  - Free online collaboratively constructed encyclopedia
  - Articles, links, categories
  - (Zesch, Gurevych & Mühlhäuser, 2007)
- Wiktionary
  - Free online collaboratively constructed dictionary
  - Words, categories, semantic relations
  - http://www.ukp.tu-darmstadt.de/software/WikipediaAPI

Textual Representation

- Concepts: Article Titles, Entry Titles, Synsets
- Article Text, Entry Information, Pseudo Glosses

Lexical Semantic Knowledge

- Semantic relatedness (SR) as measure for document relevance
- Path length (PL)
- Pseudo glosses based (Gurevych, 2005)
- Information content based
  - Resnik (1995)
  - Jiang & Conrath (1997)
  - Lin (1998)
- ESA - Explicit semantic analysis (Gabrilovich & Markovitch, 2007)
ESA: Words are Represented as Concept Vectors

In some countries, taxicabs are commonly yellow. This practice began in Chicago, where taxi entrepreneur John Hertz painted his taxis yellow based on a University of Chicago study alleging that yellow is the color most easily seen at a distance.

Computing Similarity

\[ \mathbf{v}_{\text{taxicab}} \times \mathbf{v}_{\text{truck}} = \text{Semantic Relatedness} \]

Experiments in Information Retrieval

"On the other hand, I prefer working with computers, I can program in C, Python and VB and I could therefore imagine working in the software industry."

- Topics - 30 essays of human subjects about professional interests
- Queries:
  - Nouns, Verbs, Adjectives
  - Nouns
  - Keywords (set of 41 keywords)

Document Collection

- Provided by the German Federal Labour Office
- Descriptions of 4,000 professions and 1,800 vocational trainings
- Prepared by professionals
- Evaluation on 529 descriptions of vocational trainings
- Using parts which describe profession itself, but not training or administrative details
"Gold Standard"

- 41 keywords in 3 categories
- Ranked list of professions for each topic
- Automatically extracted from knowledge base
- Used for creating relevance judgments

Relevance Judgments

41 Keywords

- educate, use/program computer,
  office, outside, animals/plants, ...

Human Annotation

Profession 1

Profession 2

Profession 3

Scoring

1. irrelevant

2. relevant

3. relevant

Outline

- Introduction: eLearning and NLP
- Automatic generation of exercises
- Assessment of learner generated discourse
- Reading and writing assistance
- Web 2.0 and computer supported collaborative learning
- Example e-NLP application: electronic career guidance
- Example e-NLP application: educational QA
- Wrap up and questions

Results

- Semantic methods lead to up to 40% improvement of search results
- Comparison of the contributions of different resources
  - Wikipedia scores best

Mean Average Precision
Question Answering (QA) vs. Information Retrieval (IR)

**INPUT:**
- Natural language **questions** and not keyword-based queries:
  - QA: How long do polar bears live?
  - IR: polar bears life span

**OUTPUT:**
- Precise and concise **answers**, not whole documents
  - QA: In the wild, **polar bears live an average of 15 to 18 years**, although biologists have tagged a few bears in their early 30s. In captivity, they may live until their mid- to late 30s. One zoo bear in London lived to be 41.
  - IR:
    - www.starbus.com/polarbear/aboutpb.htm
    - www.polarbearsinternational.org/faq/
Architecture of an Educational QA System (Gurevych et al., 2009)

Question Answering as Reuse

Question and Answer Repositories

Low Quality Questions

Question and Answer Repositories

FAQs
Ask-an-expert Services
Social Q&A sites

- Questions and answers are compiled and subject to editorial control
- Examples: www.faqs.org

- Provide expert answers to user questions
- Example: www.madsci.org

- Provide portals where users can ask their own questions and answer questions from other users
- Examples: Yahoo! Answers, WikiAnswers
Example question in Yahoo! Answers

“[YA is] the next generation of search… [It] is a kind of collective brain – a searchable database of everything everyone knows. It’s a culture of generosity. The fundamental belief is that everyone knows something”
Eckart Walther (Yahoo research)

Properties of Social Q&A Sites

- Managed by the internet community, users can:
  - Ask their own questions
  - Answer questions from other systems
- Ratings as community mechanism:
  - Points for answers, “Best Answer”, oder “thumbs up”
  - Minus points for asking a question
- The American version of Yahoo! Answers is the second-most visited education/reference site on the Internet after Wikipedia (according to Comscore)

Question Paraphrase Identification

(Bernhard & Gurevych, 2008)
Question Paraphrase Identification

### Pre-processing:
- stemming
- lemmatisation
- spelling correction

### Input question

### Question similarity measures

#### String similarity:
- matching coefficient
- overlap coefficient
- edit distance
- ngram overlap

#### Vector space:
- term vector similarity
- Lucene

### Social Q&A sites

### Results

- Vector-space based methods outperform string similarity
- Morphological pre-processing and spelling correction do not ameliorate the results

### Challenges in Question Paraphrase Identification in Social Q&A Sites

- **Spelling errors:**
  - How do you become an anesthesiologist?
  - How many years of medical school do you need to be an anesthesiologist?
- **Vocabulary mismatch:**
  - What events occurred in 1919?
  - What important events happened in 1919?
- **Solutions:**
  - Named entity recognition to identify important tokens in questions
  - Semantic relatedness metrics

### Outline

- Introduction: eLearning and NLP
- Automatic generation of exercises
- Assessment of learner generated discourse
- Reading and writing assistance
- Web 2.0 and computer supported collaborative learning
- Example e-NLP application: electronic career guidance
- Example e-NLP application: educational QA
- Wrap up and questions
NLP has lots to offer

- **Resources:**
  - Lexical semantic resources, e.g. WordNet
  - Web 2.0 resources, e.g. Wikipedia, Wiktionary

- **Tools:**
  - Tokeniser and sentence splitting
  - Morphological analysis
  - Part of speech tagging
  - Parsing and chunking
  - Word sense disambiguation
  - Summarisation
  - Keyword extraction

Tasks and applications

- To assist instructors
  - Automatic generation of questions and exercises
  - Assessment of learner-generated discourse

- To assist learners
  - Reading and writing assistance
  - Electronic career guidance
  - Educational question answering

- For all users in the Web 2.0
  - NLP for wikis
  - Quality assessment of user generated contents

What the tutorial has not covered...

- A lot more research is done on:
  - Computer-Assisted Language Learning
  - Intelligent Tutoring Systems
  - Information search for eLearning
  - Educational blogging
  - Annotations and social tagging
  - Analysing collaborative learning processes automatically
  - Learners' corpora and resources
  - eLearning standards, e.g. SCORM

NLP meets educational computing

- Educational applications are challenging for NLP since they place strong quality and robustness requirements on applications

- Interdisciplinary approach:
  - psychology
  - educational computing
  - NLP
  - cognitive and learning sciences

- Emerging types of discourse and learning paradigms in Web 2.0
How to Promote e-NLP?

- Establish an international community
- ACL and AIED associated meeting series
- Related tutorials
- Resources:
  - Bibliography
  - Research groups
  - Projects
  - Annotated corpora
  - Tools

Thank you!

http://www.ukp.tu-darmstadt.de
References

Automatic Generation of Exercises

— Computer-based Testing and Question Generation —


— Multiple-choice Questions —


— Fill-in-the-blank Questions —


— Multiple-choice Cloze Questions —


Matching Test Items


Error Correction Questions


Item Analysis


Assessment of Learner-Generated Discourse

Essay Scoring


— Plagiarism —


— Short Answer Assessment —


— Reading and Writing Assistance —

— Text Readability —


— Document Retrieval for Reading Practice —


— Text Simplification —


--- Vocabulary Assistance ---


--- Spell Checking ---


--- Grammar Checking ---


— Dictionary Lookup —


Web 2.0 and Computer Supported Collaborative Learning

— Web 2.0 —


— NLP for Wikis —


— Quality of User-Generated Content —


Electronic Career Guidance


Educational Question Answering

Bernhard, D. and Gurevych, I. (2008). Answering Learners’ Questions by Retrievi...