What Is Grammatical Error Detection?

We will be talking about errors made by learners in a second language acquisition context.

Language Learners often make non-native-like mistakes when constructing sentences:
- We arrived at the station.
- There is the garden in my house.
- I eat rice, nikujaga and salada yesterday.

Grammatical error detection entails trying to find these mistakes automatically.

Recommended reading: Leacock et al 2014, *Automated Grammatical Error Detection for Language Learners*

Where Error Detection Fits In

(a bit of a review)

- CALL: Computer Assisted Language Learning
  - Using computers and media in language learning and teaching
  - e.g., Rosetta Stone, eLanguage
  - Exercises are typically very simple in design, and offer little feedback
- ICALL: Intelligent Computer Assisted Language Learning
  - Utilize computational linguistics tools, such as POS tagging & parsing along with statistical language modeling strategies (e.g., n-grams)
    - These tools often need to be altered to expect and diagnose errors, or at least handle learner data better
    - We can also build software for specific kinds of errors. *(today's discussion)*
  - Focus on precision; we don't want to tell a learner that they've made a mistake when they haven't!

Learner Errors

Learners typically make different kinds of mistakes than native speakers.
- Content Word Choice (19.9% of all errors in CLC)
  - We need to deliver the merchandise on a daily *base/basis.*
- Preposition Error (13.4%)
  - Our society is developing *in/at high speed.*
- Determiner Error (11.7%)
  - There is *the/a garden in my house.

CLC = Cambridge Learner Corpus
Choosing the correct preposition can be a tough task even for native speakers.

There has been concern over syncing phone contacts with Facebook. “As long as you are aware of who is in the group it can be a great privacy tool. If it gets out of hand it could give you a sense of false security.” The roll out of new products comes amid reports that a syncing feature on the iPhone lets Facebook access contact data and share it on the site. “It's very possible that your private phone numbers - and those lots of your and their friends - are on the site,” said Charles Arthur. 

Choosing the correct preposition can be a tough task even for native speakers.

There has been concern over syncing phone contacts with Facebook. "As long as you are aware of who is in the group it can be a great privacy tool. If it gets out of hand it could give you a sense of false security." The roll out of new products comes amid reports that a syncing feature on the iPhone lets Facebook access contact data and share it on the site. "It's very possible that your private phone numbers - and those lots of your and their friends - are on the site," said Charles Arthur of the Guardian newspaper.

Negative Transfer: in the garden, at home, on (the) campus: same preposition in Arabic

Adjuncts: on the beach vs. at the beach

Arguments: The loaded the hay on the wagon vs. The loaded the wagon with hay

Phrasal Verbs: add up the numbers, add the numbers up

Idioms: on the house

PP Attachment: I put the ring on the table in the safe

Lexical Ambiguity: eat with a fork, view with anxiety, strike with fear, combine with others, furnish with supplies

See section 3.3.1 of Leacock et al (2014)

Language Model

Gamon et al. (2008) - Build n-grams of POS and/or parsing labels from native text and check if learner n-grams align with that model

Web-based methods - Gamon and Leacock (2010)

Take a few words of context on either side of a preposition to generate a web query
Replace the preposition with neighbors from a confusion set and search those queries
The search with the greatest number of hits is selected as the right answer


Write linguistic rules designed to find errors in learner data

Statistical - Tetreault and Chodorow (2008)

Statistical methods means building a classifier
So, what is a classifier?

Machine Learning

Machine Learning: give examples to a computer system & have it learn what the patterns are

We will explore this topic in more detail when we get to the document classification unit

Example: based on your previous purchases, what coupons should you receive?

Understanding Classifiers

Machine learning is not as scary as it sounds!

There are a number of algorithms for classification that we could talk about

Maximum Entropy, Support Vector Machines, Memory Based Learning
Each method requires different representations of information
These slides are indicative of Memory Based Learning

TIMBL http://ilk.uvt.nl/timbl/ - Daelemans et al. (2007)

Easy-ish to install
Easy to use
Works well with language data
### Running TiMBL

- We will need two sets of data:
  - Training Set - needs to be big
  - Testing Set - usually smaller
- The data sets are full of events (instances) that contain features that describe the circumstances of the event and a class that is the answer we are trying to guess
- With a little bit of a language like python, you can extract features pretty easily to train and run TiMBL
  - Open a file (e.g. a POS tagged file)
  - Extract bits of text (features) that you deem useful
  - Print those bits of text on a single line for each instance
- The real trick is selecting appropriate features

### Selecting Features

Let’s consider a real-world example:

- **The Task:** We want to classify the weather as either **good** or **bad**.
- **We would want features like**
  - temperature
  - sunny?
  - cloudy?
  - windy?
  - humidity level
  - rain/snow/none

### Machine Learning for Prepositions

- Tetreault and Chodorow used a maximum entropy classifier to try to find preposition confusions and extraneous uses
- They extracted 25 features including:
  - words/POS tags in a 2 word window(+/-) around preposition
  - the head verb and noun of the preceding VP and NP
  - the head noun of the following NP
- **John went to the store this morning.**
  - word+POS bigrams: went_VBD, the_DET
  - head of previous VP = went
  - head of previous NP = John
- Their system achieved 84% precision and 19% recall.
  - This might sound low, but keep in mind, we want to get the best possible precision, even if it means losing recall.

### Types of systems

**Summary**

Systems differ in terms of:

1. the kinds of features they use
   - surface level features, syntactic features, L1 information, etc.
2. the training data they use
   - correct usage, artificially generated errors, real errors
3. the kinds of models (e.g., classifiers) they use
   - classifiers, language models, web counts, etc.