Reading Notes

- Sarcasm Detection
  - Detecting Sarcasm is Extremely Easy ;) (Parde & Nielson 2018)
  - Harnessing Context Incongruity for Sarcasm Detection (Joshi et al 2015)
  - Sarcasm as Contrast between a Positive Sentiment and Negative Sentiment
- Neural Networks
  - Catastrophic Interference in Neural Embedding Models (Dachapally & Jones)
  - Querying word embeddings for word similarity and relatdness
  - Multi-Task Deep Neural Networks for Natural Language Understanding
- Answer Scoring
  - Riordan et al., 2019
  - Horbach et al., 2019
  - Riordan et al. 2020
- CLINGDINGS
  - How do you determine the worth of a language?
  - November 6th 2019: Hai, Peng
  - Alan Ridel
  - Hai Hu 02-19-2020
  - Zeeshan 02-19-2020
- Parsing
  - Overview of the SPMRL 2013 Shared Task:Cross-Framework Evaluation of Parsing Morphologically Rich Languages
  - Dependency Parsing
  - Characterizing the Errors of Data-Driven Dependency Parsing Models
- Reading Template
- Professionalization workshop
  - January 17th - Job search
  - Job talk Monica Nesbit
- Language Modelling

    - BLiMP: A Benchmark of Linguistic Minimal Pairs for English
- Swahili Syntax
    - Swahili Syntax (Anthony Vitale, 1981)
- Syntax for "Exotic" languages
    - Developing Universal Dependencies for Wolof
    - Towards a dependency-annotated treebank for Bambara (Aplonova & Tyers 2018)
- To Read
- Universal Dependencies
    - A Universal Part-of-Speech Tagset (Petrov, Das, McDonald)
    - Universal Depedencies v1: A Multilingual Treebank Collection
- CG to Dependency Parse
    - Reusing Grammatical Resources for New Languages
    - Estonian Dependency Treebank: from Constraint Grammar Tagset to Universal Dependencies
- Bantu NLP
    - Learning Morphosyntactic analyzers from the bible via iterative annotation projection across 26 languages

Sarcasm Detection

Detecting Sarcasm is Extremely Easy ;) (Parde & Nielson 2018)

Gist

- Doain general sarcasm detection system
- Applied to twitter and amazon product reviews
- Contains error breakdown

Intro

- Sarcasm is difficult even for humans
  - Primariy indicated using prosodic rather than syntactic cues
- Previous approaches have been largely domain specific, this is an attempt at a general purpose sarcasm detection system

Background

- Tweets may be expecially challenging because the text limit may encourage brief coments that require more contextual information
  - The example of saying "Great" just after an election may be understandable to others at that point in time but for an automatic system that is not aware of such events, it

becomes very difficult.

- Rajadesingan et al 2015 "developed behavioral models of sarcasm usage specific to individual users" (p. 22)
- Sarcastic tweets are sampled using hashtags indicating sarcasm, Amazon reviews are sampled using star ratings
- The prior work (Parde and Nielson 2017) created a domain adaption system that was used prior to training the model, this achieved better performance "in predicting sarcasm in Amazon product reviews over models that trained on reviews alone or on a a simple combination of reviews and tweets" (p. 22)

Sarcasm detection methods

Data source

- Train
  - 3998 tweets, 1003 Amazon product reviews
- Test
  - 1000 tweets (609 non-sarcastic and 391 sarcastic)
  - 251 amazon reviews (87 sarcastic and 164 non-sarcastic)

Features

- Contains Twitter Indicator
  - "Multiple binary features indicating whether the instance contains one of th esarcasm-related has-tags, emoticons, and/or indicator phrases learned by Maynard and Greenwood (2014)" (p 23)
- "Twitter-Based predicates and situations
  - "Multiple binary features indicating whether the instance contains a positive predicate, a positive sentiment and/or negative situation phrase learned by Riloff et al. (2013)

from a corpus of tweets. Includes an additional binary feature that indicates whether one ofo those positive preedicates or sentiments precedes one of those negative situation phrases by <= 5 tokens"

- Star Rating
  - "Number of stars associated with the review" (p 23) left blank for tweets
- Laughter and interjections
  - "Multiple binary features indicatingi whether the instance contains: hahahaa, haha, hehehe, hehe,jajaja, jaja, lol, lmao, rofl, wow, ugh, and/or huh" (p 23)
- Specific characters
  - "Multiple binary features indicating whether the instance contains an ellipsis, an exclamation mark and/or a question mark" (p 23)
- Polarity
  - "Multiple features indicating the most polar (positive or negative) unigram in the instance, the polarity score (-5 to +5) associated with that unigram, the average polarity of the instance, the overall (sum) polarity for the instance, the largest difference in polarity between any two words in the instance, and the percentages of positive and negative words in the instance" (p 23)
- Subjectivity
  - "The percentages of strongly subjective positive words, strongly subjective negative words, weakly subjective positive words, and weakly subjective negative words in the instance" (p. 23)
- PMI
  - "Multiple features indicating the highest number of consecutive repeated characters in the instance (e.g., Sooooo => 5) and the higehest number of consecutive punctuation characters in the instance" (p 23)
- All-Caps
  - "Multiple features indicating the number and percentage of all-caps words in the instance" (p. 23)
- Bag of words
  - Features for words most closely associated with the different training pairs (e.g. Amazon - Sarcastic, Amazon non-sarcastic, twitter sarcastic etc.)
  - Features for most common words in each of these different class source pairings.

Classification Algorithm

Naive bayes using Daume III (2007)'s method for domain adaptation. to generate source, target and general feature mappings.

Results

.59 F-score on twitter data, 1% over previous literature (not really meaningful) Recall of system is much higher (.68 vs .62) at the cost of some precision (53 vs 55). .78 F-score on Amazon reviews, much higher than previous results (Buschmeier et al 2014) (78 to 74). Once again, much higher recall (82 to 69) at the cost of precision (75 to 85)

Error analysis

- Many did not convey sarcasm once the sarcastic hash tags were removed (23)
- 8 only had sarcastic content in the hashtags
- 13 tweets were discovered not to be sarcastic upon manual inspection
- 63 Required world knowledge to know that it was sarcastic.
- Highly negative
- Reviews also had story-like passages that were sarcastic. E.g. a narrative where the thing being reviewed is doing things that are impossible.

Sarcasm Detection

Harnessing Context Incongruity for Sarcasm Detection (Joshi et al 2015)

Gist

- The key part of this paper is that incongruity e.g. clashes in sentiment are central to the detection of sarcasm
- "It must be noted that our system only handles incongruity between the text and common world knowledge (i.e. the knowledge that '*being stranded*' is an undesirable situation and, hence, '*Being stranded in traffic is the best way to start my week*' is a sarcastic statement)." (p 758)
- "This leaves out an example like '*Wow! You are so punctual*' which may be sarcastic depending on situational context" (p 758)
- Explicit Incongruity is where there are polarity signifying words that make the clash in sentiment apparent
- Implicit incongruity is where there are phrases that imply a particular sentiment conventionally. **These are the ones that seem the most interesting to see how they deal with them**.

Dataset

Primarily focused on tweets.

- Tweet-A (5208 Tweets, 4170 sarcastic) Downloaded by looking for certain hash tags (#sarcasm, #sarcastic adn #notsarcastic) and then did a rough quality control check to make sure that they made sense, removing wrongly labeled examples.
- Tweet-B (2278 tweets, 506 sarcastic) manually labeled for Riloff et.al 2013. I suspect what they're doing here is trying to balance the class distributions for this since predicting sarcastic tweets using the Tweet-B dataset would be quite difficult.

Discussion board datasets

- Discussion-A (1502 discussion board posts, 752 sarcastic). Obtained from the Internet Argument Corpus (Walker et al. 2012). Manually annotated,. 752 sarc and non-sarc posts are selected randomly.

ML System

Detecting incongeruity

- Identifying phrases with implicity sentiment
- Obtained using algorithm given in Riloff et al. (2013) but extract both possible polarities for both nouns and verbs
- Keeping subsumed phrases "(i.e. `being ignored' subsumes 'being ignored by a friend')"
- Riloff et al. 2013 used these phrase as part of rules while this approach is a ML approach that uses them as features.

Features

- Unigrams
- Number of capital letters
- Number of emoticons and lol's
- Number of Punctuation marks
- Boolean feature indicating whether implicitly incongruous phrases were extracted.

Explicit Incongruity features

"""

- Number of times a word is followed by a word of opposing polarity
- Length of largest series of words with polarity unchanged
- Number of positive words
- Number of negative words
- Polarity of tweet based on words present """

Analysis

- Ran into errors with subjective things (Maybe this would be resolved if they wre able to look more closely at a user's history)
- Errors when there was incongruity but it was not within the text
- Incongruity due to numbers causes errors, here's the example they provide "*going in to work for 2 hours was totally worth the 35 minute drive*"
- Pieces of sarcastic text embedded in a larger non-sarcastic text were harder to identify.
- Politeness of sarcasm introduced difficulties.

Sarcasm as Contrast between a Positive Sentiment and Negative Sentiment

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, Ruihong Huang

Novel bootstrapping algorithm that learns lists of positive sentiment phrases and

> "Bootstrapping algorithm that automatically learns phrases corresponding to negative sentiments and phrases corresponding to negative situations" p. 705

Bootstrapped learning of positive sentiments and negative situations

> "Our goal is to create a sarcasm classifier for tweets taht explicitly recognizes contexts that contain a positive sentiment contrasted with a negative situation" p. 706

They're learning phrases that have positive or negative connotations using a single seed word "love" and a collection of sarcastic tweets.

> "Operates on the assmption that many sarcastic tweets contain both a positive sentiment and a negative situation in close proximity, which is the source of the sarcasm" p. 706.

They focus on positive verb phrases and negative complements to that verb phrase.

They don't parse because, well, parsing tweets is messy and hard. Instead they use just part of speech tags and proximity as a proxy for syntactic structure.

> "We harvest the n-grams that follow the word 'love' as negative situation candidates. WE select the best candidates using a scoring metric and add them to a lsit of negative situation phrases. p.706

> Next we explait the structural assumption in the opposite direction. Given a sarcastic tweet that contains a negative situation phrase, we infer tha tthe negative situation phrase is preceded by a positive sentiment. We harves the n-grams that preceed the negative situation phrases as postive sentiment candidates, score and select the best candidates, and add them to the list of positive sentiment phrases" (p. 706)

Using only 175,000 tweets... Quite small for such distantly supervised stuff to work.

They use #sarcasm as indicative of the sarcastic class.

They use part of speech patterns to identify verb phrases and noun phrase.

They're scoring each candidate based upon how well they corresond with sarcasm. E.g. "we score each candidate sentiment verb phrase by estimating the probability that a tweet is sarcastic given that it contains the candidate p hrase preceeding a negative verb phrase" p. 708

and "we score each remaining candidate by estimating the probability that a tweet is sarcastic given that it contaisn the predicative expression near (within 5 words) of a negative situation phrase"

> We found that the diversity of positive sentiment verb phrases and predicative expressions is much lower than the diversity of negative situation phrases

Makes good sense that they found this ^ However, they seem to have more stringent filtering for the positive expressions...

Neural Networks

Catastrophic Interference in Neural Embedding Models (Dachapally & Jones)

Catastrophic forgetting is the tendancy of neural models to have a strong recency bias e.g. more recent training examples are more likely to be predicted.

DSM

Distributional Semantic Models encompass geometric models like latent dirchlet allocation and svd as well as neural embedding models. Neural embedding models are

Experiment 1

Create artificial data

using the following sentence generation patterns

- "Man/woman catch/eat trout/bass"
- "Man/woman play/pluck acoustic/bass"

The idea is to capture the two homophonous meanings of 'bass' and place them in embedding contexts identical to that of a synonym.

Ordering of data

Balancing distribution of homophones

- Random sampling
- All 'fish' interpretation s first
- All 'musical' interpretations first

1/3 of one meaning

Evaluation

Looked at cosine similarity between word embedding vectors learned

Experiment 2

Conducted using real data TASSA corpus

Neural Networks

Querying word embeddings for word similarity and relatdness

Word relatedness is sometimes asymmetrical e.g. stork may elicit associations with baby but baby may not generate associations with stork.

Similarity is symmetrical.

Neural Networks

Multi-Task Deep Neural Networks for Natural Language Understanding

Answer Scoring

Riordan et al., 2019

How to account for mispellings:

Quantifying the benefit of character representations in neural content scoring models

Takehome:

"Models with character representations outperformed their word-only counterparts...lower MSE and higher QWK" p. 121

Datasets

- ASAP-SAS: 10 questions with large number of responses for each question, sentence or two in length

- Formative-SAS: dataset collected by ETS (relatively short answers)
- Summative-LAS: 20 questions, mean number of words is 230

Methods

Word only model

pretrained word embeddings into bidirectional GRU. Hidden states of GRUs are either pooled or go through an MLP attention mechanism Output of the encoder goes through sigmoid fully connected layer which produces a score

Character + word models

Each word is represented with a sequence of 25-dimensional character embeddings. "Character embeddings are concatenated with the word embeddings prior to the word-level encoder" (p. 119)

Results

ASAP-SAS

While adding character representations performed better than just spelling correction, the effect of adding character representations was not statistically significant in the GLMM model and using spelling corrections was not significant either.

No evidence for interaction between character representations and spelling correction in the GLMM.

Formative K12-SAS

Same general trend as ASAP-SAS

- character and word representations outperform word representations
- spelling corrected models outperformed non-spelling corrected models

Statistical significance between the different representations and the different methods of spelling correction but no interaction observed between mispelling bins and the representation used.

"The difference between feature sets and between mispellings bins was significant even when controlling for score and number of words" (p. 123)

Large majority of responses had no spelling errors. 3 spelling bins used (0, 1, 2+)

Q: Is spelling not what the character representations are able to capture? Is it instead morphological variation?

- What if you ran a stemmer over the input? Would the difference between word+character embeddings and plain word embeddings go away? Surely someone has done this.

Q: I thought that the addition of character representations was helpful for two of the datasets but not the last one. The conclusion reached was that character representations were not as helpful as spelling correction but I think this was only significant for the 2nd dataset.

Q: Are the character representations alone enough? (what if you dropped words)

Answer Scoring

Horbach et al., 2019

The influence of variance in learner answers on automatic content scoring

Andrea Horbach and Torsten Zesch

Variance

Sources of variance

- Conceptual variance:
    - when there are multiple separate right answers to a question.
    - bigger issue is number of variants of incorrect answers. *why not focus on modelling correct answers? Could you use an approach that allows you to rely more on how close this answer is to the correct answers I saw in training (if generative, I'm not sure how this would work for discriminative) could you model correct/wrong questions as anomaly detection?*
- Variance in realization
    - different ways of forming the same conceptual answer

- Linguistic variation
    - language provides lots of possibilities to express the same meaning *what if you did reparsing or something to map variant forms to roughly the same meaning*

Riordan et al. 2020

An empirical investigation of neural methods for content scoring of science explanations

NGSS science standards dimensions

- DCI (disciplinary core ideas
- CCC (cross cutting concepts)
- SEP (science and engineering practices

KI rubric:

- > involves a process of building on and strengthening science understanding by. incorporating new ideas and sorting out alternative

> perspectives using evidence

- rewards conecting evidence to claims in their explanations

Data

Constructed reponse (CR) items are evaluated. The ones chosen are cases where SEPs need to be used while showing understanding of CCCs and DCIs.

CR Items:

- Musical Instruments and the Physics of Sound Waves (MI)
- Photosynthesis and Cellular Respiration (PS)
- Solar Ovens (SO)
- Thermodynamics Challenge (TC)

Two separate rubrics in parallel:

- KI rubric
    - linkage with subsets of the ideas described in the evidence statements
        - Photosynthesis (PS) listed 5 ideas related to energy and matter changes during photosynthesis
    - Scores from 1-5
- NGSS subscore rubric
    - two of three dimensions for each CR
        - Only those that are relevant given the prompt are used (e.g. a question where the answer doesn't depend upon science and engineering practices would not have a score for that dimension)
    - scores from 1-3

The thermodynamics challenge item was particularly challenging.

Sometimes there were less annotated data available for the NGSS dimension models compared to the KI models.

Models

Each item and score type were trained independently. 10-fold cross validation with train/val/test splits, evaluating on concastenated predictions across folds.

SVR

- binary word unigrams and bigrams

RNN

- pretrained word embeddings (GloVe 100) fed into a bidirecitonal GRU encoder.
- Hidden states of GRU are pooled (max)
- Encoder output aggregated in a fuly-connected feedforward layer using sigmoid act (giving scalar score).
- **Presumably the same scaling and unscaling is happening that we worked with before because sigmoid should be squishing everything to be between 0,1**
- exponential moving average across weights used during training
- 50 epochs

Pretrained transformer

- bert-base-uncased
- using `[CLS]` token output, fed through a non-linear layer to obtain the scalar score.

- exponential moving average across weights used during training
- 20 epochs
- When identifying best hyperparameters, for each fold, taking the epoch where validation performance is highest for evaluation.
- During final training, validation and training data are concatenated and then the model is retrained.
    - **I assume this is done for all the models but it's only mentioned for the PT model**

Results

KI models

The Pretrained transformer models are more robust, they're always ahead of the RNN on all metrics (sometimes not by much though).

The items that were highly skewed showed lower levels of human-machine agreement (lower thant he 0.7 threshold for QWK in real world scoring applicaitons) **Where does that threshold come from??**

CLINGDINGS

CLINGDINGS

How do you determine the worth of a language?

How do you determine the worth of a language?

Arle Lommen October 30 2019

2019 is the UN year of indigenous languages.

- Highly idealized language ideals
  - Everyone be able to use their own language as they see fit.
  - Obviously this isn't exactly how this works.

Every language has an intrinsic value

However, in a world of limited

resources, not all 7,000 of the world's languages can be invested in.

less than 1% of content is translated into another language

To cover 100% of content this would take about 20 million translators. This is only for one additional language though.

to cover all 135 economically important languages we would need 2 billion translators.

Let's look at other views of value

the number of speakers does not determine the value of a language for a business

- Though this may be exactly the information that is relevant to NGO's or religious organizations.

Maybe we can look at GDP?

- Could look at GDP per capita to get a sense of the wealth of the individual **I'm not sure why wealth per individual is

Internet adoption rate

- More relevant to today's globally connected tech powered companies.

Pre 2019 CSA was selecting 50? language with online relevance (e.g. usage by communities online).

Calculated number of speakers for each country/territory.

Used a zero-sum approach (no accounting for multi-lingualism).

Assigned languages to four tiers basd on cumulative market research.

This measure is called eGDP or electronic GDP, this is not a measure of ecommerce.

after 2019,

- Added multilingualism
- they expanded from 300 locales to 500 locales.
- Added model of income inequality to help scale GDP (e.g. if 12% of a country's pop is online, those people probably have higher GDP than the average of the whole population)

November 6th 2019: Hai, Peng

Product harm report evaluation

Product harm crises are when products cause indicents and lead to issues and then the public response produces negative publicity for the company and the government body in charge of regulation

Two issues for issuing a recall

- Delayed announcement of recall
  - Food recalls take an average fo 57 days after discovery
  - Automotive recalls average of 306 days later (US))
- Low recall completion (small proportion of products that should be recalled are recalled)

Legal wiggle room

- Have to explain how recall was discovered
- What steps were taken to determine whether recall should be done
- free to determine how they release the information and how much information they release

Research questions

do recall commmunication examples differ across industries

Hypothesis

- LOnger the recall takes the worse the company is viewed
- The more steps taken the more optimistic the more favorably the company is viewed

the idea is that these shape the way that the company frames their response.

- The model needs to account for year effects, firm effects, etc.
- dependent variable is linguistic variables
- independent variable is number of steps taken by the company and time taken to report

argument structure is crucial for previous

research, in addition to subjectivity measures

difference emerges in number of content words (nouns, verbs adjectives and adverbs)

word (lexical complexity)

- MATTR (moving average ttr)
- STTR mean ttr for every 100 words
- CTTr (corrected ttr) types / sqrt(2 * #tokens)

Structural complexity (length of t-unit + dependency length (what is

a tunit?)

Reading ease score takes into consideration number of syllables per word and number of words per sentence.

However, the number of syllables per word is hard to reliably calculate.

CLINGDINGS

Alan Ridel

We know that there were a large number (25,000 books published during the victorian era) of books, we have a lot of information about gender and year level stats.

no corpus that exists reflects the population of published novels during this period effectively.

The Chadwyck-Healey corpus is particularly bad, 50% of the data comes from male authors published before 1876 even though this was only 15% of the population.

Random sampling of the population is not really possible because we don't actually have a complete database of all novels published during the victorian era.

instead we do quota sampling.

We divide up the population into categories based on year and gender and manually encode a randomly selected chapter.

- Not a representative sample
    - overrepresents authors who wrote more than one novel
    - over represents novels published in multiple volumes

Maybe there's a bias in which things were published or which types of genres tend to do multi volume things

The solution is to use post-stratification as a way to do analysis of granular distinctions after the fact:

- e.g. novels published by women in 1940
- novels involving trains

CLINGDINGS

Hai Hu 02-19-2020

Building a natural language inference dataset in Chinese

What is NLI?

when you have to determine whether a hypothesis contradicts, entails from or is neutral towards a premise.

Issues with SNLI

Turkers do not want contradiction to go both ways.

Bias in hypotheses

If you train on SNLI on just the hypotheses, you get better than majority baseline.

There's bias in the hypotheses One thing is that sleeps contradicts almost any other action. Additional heuristics in the dataset probably introduced by the Turkers probably exist. By creating synthetic data that goies against the heuristics, the result is very very poor performance (19% accuracy for BERT was the best).

XNLI:

- 15 languages
- translated from SNLI/MNLI
    - bad quality translation, lots of things that just don't translate well

Our chinese NLI

- undergrads instead of turkers
- told to write 3 neutral, 3 contradiction, 3 entail as a way of getting them to introduce more variety.
- Students still apply heuristics.
- Issues that emerged:
    - phone call transcriptions are bad
    - use of questions in premises was confusing

Todo

- how to get more variation in hypotheses?
- one annotator only writes Entailments not C/N

Zeeshan 02-19-2020

Internship at Amazon and forthcoming thesis

What is transfer learning?

- Transfer learning is a a variety of different things. For a taxonomy read Ruder 2019.
- pretraining of word embeddings is probably the most famous form of transfer learning.

Multi-task learning

Hard vs soft parameter sharing Hard parameter sharing literally shares some of the initial layers and then has task specific layers towards the end.

Soft parameter sharing uses soem method of regularization to force common layers for the two tasks to be close to eachother.

Parsing

Overview of the SPMRL 2013 Shared Task:Cross-Framework Evaluation of Parsing Morphologically Rich Languages

Central topic

- Provide standard datasets for morphologically rich languages in different representations and parsing scenarios.
- Standardize the evaluation protocol on morphologically ambiguous input
- Raise community awareness with regard to the difficulty of parsing morphologically rich languages

Methodology

Datasets

- Include data in both constituency and dependency annotation.

- *full data* setup and *small* setup (5,000 sentences)
- Three parsing scenarios:
  - gold segmentation, pos tags, and morphological features are provided
  - automatically predicted segmentation, pos tags and features
  - lattice of multiple possible morphological analyses and joint disambiguation of the morphological analysis and syntactic structure

Findings

Previous research

- first statistical parsing models were generative and based upon treebank grammars
- > applying the phrase-based treebank grammar tecniques is sentsitive to language and annotation properties, and these models are not easily portable across languages and schemes.

Notable quotes

> While progress on parsing English -- the main language of focus for the ACL community -- has inspired some advances on other languages, it has not, by itself, yielded high-quality parsing for other languages and domains. This holds in particular for morphologically rich languages... where important information concerning the predicate-argument structure of sentences is expressed through word formation, rather than constituent-order patterns as is the case in English and other configurational languages. p. 146

> “recently, advances in PCFG-LA parsing (Petrov et al. 2006) and language-agnostic data-driven dependency parsing (McDonald et al. 2005; Nivre et al. 2007b) have made it possible to reach high accuracy with classical feature engineering techniques in addition to, or instead of, language specific knowledge. p. 147

Follow up readings

Parsing

Dependency Parsing

Parsing

Characterizing the Errors of Data-Driven Dependency Parsing Models

McDonald & Nivre 2007

Background

Two basic approaches to dependency parsing are all pairs and stepwise.

- All pairs = graph based
- stepwise = transition based

"All pairs" approaches make decisions globally, use exact inference but have relatively impoverished features

"stepwise" approaches make greedy decisions, but have a rich feature representation including past decisions.

Both achieve similar performance but the kinds of errors they make are different. Segue and Lavie (2006) shwo that combining the predictions of both types of models yields "significantly improved accuracy" This paper is going to talk about the strengths and weaknesses of the approaches.

Two models for dependency parsing

Preliminaries

Describes what a dependency tree is, & graph based and transition based dependency parsing. Overall, Kuebler et al (2009) has a more thorough discussion of the different approaches.

Global Graph based parsing

Labels are portrayed as part of the scoring function in this work. *I believe how scoring labels works varies between different approaches but I have to look further into this*

> The primary disadvantage of these models is that the feature representation is restricted to a limited number of graphs arcs. This restriction is required so that both inference and learning are tractable

MSTparser is the implementation used.

Local, Greedy, Transition-Based Parsing

> The primary advantage of these models is that afeatures are not restricted to a limited number of graph arcs but can take into account the entire dependency graph built so far. The main disadvantage is that the greedy parsing strategy may lead to error propogation.

CONLL-X shared task

13 languages 19 systems labeled attachment score was official metric (percentage of tokens, excluding punctuation, that are assigned botht he correct head and the correct dependency label).

Error analysis

Graph factors

All current parsers have more trouble on longer sentences. MaltParser performs better in shorter sentences, worse as sentences get longer. Attributed to likelihood of error propogation being higher for longer sentences and richer feature representation as beneficial for short sentences.

MSTParser far more precise than MaltParser for longer dependency arcs (where the length is the length of the predicted arc). MaltParser does better for shorter dependency arcs. Overall MSTParser is not affected by dependency length.

MSTParser is far more precise close to the root and is less precise then Malt further from the root.

Dependency arcs further from the root are (usually) created first in transition based systems. Thus this is further evidence that error propogation is partly to blame for the difference between the two approaches.

> MSTParser over predicts arcs near the bottom of the graph. Whereas MaltParser pushes difficult parsing decisions higher in the graph, MST Parser appears to push these decisions lower

Linguistic Factors

Findings with regard to part of speech associations are tied to previous findings of position in graph.

Adpositions are a bit strange because they have high average root distance and low average dependency length but MSTParser does okay on them.

Reading Template

Central topic

Methodology

Findings

Follow up readings

Professionalization workshop

January 17th - Job search

Why this is important:

- Grad students interact directly with the faculty
- A number of professors have retired or will be retiring soon.

some facts about IU linguistics

- IU ling department is historically gender imbalanced
    - Neither female faculty members were full time in the linguistics department (in 2000)
- Lack of certain subfields
    - No semantics
    - applied linguistics was getting split off
        - under the school of education in the 1960's
        - moved under linguistics in the 1970's
        - 2006: divorce of linguistics and applied linguistics
- Currently, no one in sociolinguistics at IU
- Faculty who are still year and were here in 2000 will probably be retiring in the next 5 years

should new hires reflect the

traditional strengths of the department or cutting edge research?

- African linguistics is a particular subfield that is a historic strength but may not be an area of cutting edge research

The talk on the 26th is a sociolinguistics talk, all the others are syntax talks.

if you have thoughts then you should contact Nils or Samson

Syntax job search

- 68 applicants

Quality of candidates is important

Fit in the department is equally

important

- Trying to find someone with a secondary specialty that fits well with the department
- Specialities in languages other than English

Be careful when searching for a job

- It's a serious decision for the institution and the applicant
- academia works in much slower hiring cycles, it takes up to a year to remove someone who was just hired

After job talks the whole remainder of the afternoon is for grad students

- We'll just assemble in the seminar room
- You'll get a picture of how scholars interract with grad students
    - This is important for candidates because involvement with grad students is essential
- if you can't make it to the job talk or the after discussion, send Dr. DeJong a note and see if something can be arranged

What should we do to prep?

- What the faculty care about and what we care about is just different
- Our role is to go to every event we can. even if we're not interested in syntax because it impacts everyone.
    - Going to talks you don't understand is a great way to learn and it's also important
- Don't come into this cold, read up on their work on google scholar a bit to get an idea of what they're into
- A campus visit is like a marathon. Campus visits are grueling.
    - Be considerate of the candiates, they are humans and they probably need things like coffee, the bathroom whatever
    - Personal things are off limits for conversation
- Pretty dramatic difference applying before your dissertation is complete.

We may have a debriefing session after them?

Professionalization workshop

Job talk Monica Nesbit

Lansing speech corpus:

- 106 interviews
- Sociolinguistic interviews conducted every 2 years
    - Examining different situations
- Oral histories
    - collected during early 2000's
    - only casual conversation

Had to exclude people

- Non-white
- Rural people
- These groups do not normally show the northern city shift

none of the younger speakers show the hallmarks of the northern city shift.

is ash [+tense]?

People generally don't want to have syllables without codas for lax vowels. What do people from Lansing, Michigan do?

Younger northern inland speakers pattern more closely to the candian model.

Widespread perception that inland northern speech is 'correct'

- More recently, there have been discussions that are less positive towards the northern city shift.
- Interviewed people have some strong negative comments when they hear older Norther City speakers.
- Perception that people with northern city shift are uneducated but hard working.

Lansing: auto town

- original home of oldsmobile
- headquarters of gm for 20 years.
- many factories closed after the move from Lansing
    - Shift to service type industry
- With the shift to a service industry, blue collar workers went from being prestigous to marked
- Dialect attrition is happening across the northern inland region

Low back merger shift is a western change shift. Does the chain shift happening in California match the northern city shift? Turns out no not really? Loss of local dialects paralleled in New England as well (rsearch at Dartmuth).

Language Modelling

Language Modelling

BLiMP: A Benchmark of Linguistic Minimal Pairs for English

The paper and dataset can be found here.

Dataset

- 67 sub-datasets each containing 1000 minimal pairs isolating specific contrasts in syntax morphology or semantics.
- Automatically generated using grammars
- Used to evaluate several different types of state of the art language models
  - > identify morphological contrasts reliably but struggle with semantic restrictions on distribution of quantifiers negative polarity items and subtle syntactic phenomena such as extraction islands

Island phenomenon are the hardest for language models to deal with, scores on these minimal pairs are near chance. (which is funny because these are fairly robust restrictions in human grammacality judgements).

What if you trained a language model with negative examples using the minimal pairs provided? Frame it as a classification problem and see how they compare then? Kind of avoids the issue at *heart since the paper is looking at how well existing language models address these phenomenon but it would be interesting to see if these architectures* ***can*** *model this information*

The CoLA dataset features 10,000 judgements and that shows BERT and company doing well at that task. (this is mentioned later)

Related work

Language modelling

recent shifts to transformer models have resulted in reduced perplexity however, "this doesn't give insight into these models' linguistic knowledge."

> Evaluation on downstream task benchmarks (Wang et al. 2018, 2019a) is more informative, but might not prsent abroad enough challenge or represent grammatical distinctions at a sufficiently fine-grained level.

Evaluation of linguistic knowledge

There have been previous works that examined using minimal pairs to infer whether language models learn about specific linguistic phenomenon.

However, most of these works have been limited in what they investigated:

- LInzen et al. (2016) look closely at subject verb agreement
- Marvin and LInzen (2018) look at a larger set of stuff including NPI and reflexive licensing.
- Things like control, raising, ellipsis etc have not been included despite being well studied linguistic phenomenon.

There are corpora that contain grammaticality judgements for sentences. The most recent and largest is CoLA (WArstadt et al. 2019b). CoLA is included in the GLUE benchmark.

Current transformer models can be trained to give excellent results on this data.

looks like my previous idea has already been done

When Warstadt and Bowman (2019) investigaated the performance of pretrained language models including an LSTM, GPT and BERT, they found that the models did well on "sentences with marked argument structure" and did worse on sentences with long-distance dependencies (though transformer models did better there).

> evaluating supervised classifiers prevents making strong conclusions about the models themselves, since biases in the training data may affect the results

Performance could be due to the occurance of similar examples in the training dataset.

When language models are evaluated on minimal pairs, this evades the problem.

The authors say the probability of a sentence (and thus the inverse perplexity) can be used as a proxy of acceptability.

Dataset

The data is automatically generated using expert crafted grammars.

- ensures that there are sufficient unacceptable answers (which are very very rare in naturally occuring text)
- allows for fully controlled dataset with isolation of each linguistic phenomenon.

Data generation procedure:

- Use a basic template

- pull from a vocab of 3,000 morphologically, syntactically, and semantically annotated words
    - These features are needed to create grammatical and felicitous sentences
- Code is available in this github repo

Sometimes implausible sentences can be generated but the authors view this as a non-issue.

The authors consider frequent inclusion of a phenomenon i na syntax/semantics textbook as an informal proxy for what is core linguistic phenomenon for English. (not especially useful when examining non-English languages as few are taught from the perspective of a different language. E.g. a minimalist syntax textbook that only discusses French)

These are the phenomenon covered

- Anaphor agreement
- Argument structure
- Binding
- Control/Raising
- Determiner/Noun agreement
- Ellipsis
- Filler-Gap
- Irregular forms
- Island effects
- NPI licensing
- Quantifiers
- Subject-Verb agreement

Comparison to related resources

with 3000 words, this has the widest vocabulary of any related generated dataset. 11 different verb subcategorization frames.

Other works like Linzen et al (2016) that use a larger lexicon size but use data-creation methods that are limited in control or scope.

- Linzen (2016) change number marking on present tense verbs but this is a strategy that is specific to subject agreement phenomenon
- Lau et al. (2017) build a dataset but doing multiple round trip machine translations but this creates a number of grammatical violations and does not offer the granular minimal pairs that this paper's data generation method provides.

Validation

Used mechanical turk 20 annotators rated 5 pairs from each of the 67 paradigms. aggregate human agreement is estimated at 96.4%.

Only cases where the annotators agreed with Blimp on 4/5 examples from each paradigm were included. (the 67 paradigms included passed, 2 additional ones were rejected on these grounds).

Individual human agreement approximated at 88.6%

Evaluation of language models

GPT-2 achieves highest scores, n-gram the lowest, LSTM and Transformer-XL tied.

> the results seem to indicate that access to training data is the main driver of performance on Blimp for the neural models we evaluate

They point to the fact that the LSTM and Tranformer-XL models performed about the same despite wildly different architectures and GPT-2 had 100x the training data but similar architecture to Transformer-XL.

Phenomenon specific results

models perform best to human level on morphological phenomena (anaphor agreement, determiner-noun agreement, and subject-verb agreement). *Possibly because english doesn't have that much of this*

GPT2 is the only model that performs above chance on Islands but it is still 20 points behind humans. they are very hard in general.

Wilcox et al (2018) concluded that LSTMs have knowledge of some island conditions which contridicts the findings here. However, Wilcox et al. compare four related sentences with or without gaps, obtaining wh-licensing as a metric of hos strongly the language model identifies filler-gap dependency in a single spot, the lm has learned the constraint if the probability is close to 0. *this is difficult to parse, I think I need to read the original paper*

This paper finds that neural models can identify long-distance dependencies but not the domains where these dependencies are blocked.

weak performance on argument structure is somewhat strange because previous work has suggested that argument structure is a solid domain for neural models. However, these works (Warstadt and Bowman (2019)) trained the model on CoLA and didn't do direct language modelling.

Contribution

I am unfamiliar with the creation of minimal pair datasets for evaluation of neural language models. It seems that this paper's main contribution, though, is the creation of their new dataset that approaches minimal pairs with more breadth: including examples of many more types of

English linguistic phenomena. They have the widest vocabulary of any generated dataset like this, including a large number of verb subcategorization frames.

Swahili Syntax

Swahili Syntax (Anthony Vitale, 1981)

Grammatical Sketch

Very brief grammatical sketch, strong focus on syntax which is nice to see since most grammar sketches avoid syntax as much as possible :)

SVO structure

> Swahili is a positional language rather than a case language. That is, it is at least partly the position of constituents in a phrase-marker which determines grammatical relations such as ''subject'', ''object'' and so on. (p. 18)

- Proposes SVO as the canonical word order with the subject defined at the NP directly dominated by S and the object being the NP daughter to V. (no internal subject hypothesis here).

Variations in word order

> Word order may differ from the normal SVO sequence due to such factors as

> emphasis, definiteness, and type of information (i.e. "old" vs "new"). (p 19).

Permutations are typically unambiguous due to the very clear verbal morphology indicating the noun class + person/number of the subject and object.

> “If both NP's contain the same feature specifications for class and person, a late movement rule such as this one is typically blocked (p. 19).

These are typically interpreted as having an SVO order unless intonational changes accompany the sentence (see Maw 1969).

Simplex sentences

Complex sentences

The syntax of voice

Theoretical implications

Going to skip this part because I'm working in a completely different framework (and, in fact, most generative syntacticans are as well).

Syntax for "Exotic" languages

Syntax for "Exotic" languages

Developing Universal Dependencies for Wolof

The paper can be found here.

Wolof is a Niger-Congo language (however it is Senegambian where Bantu languages are Benue-Congo (citation needed)).

Computational grammar of Wolof in the LFG framework (page 1). This was used to create the first treebank of Wolof (see the ParGramBank paper Sulger et al. 2013).

The dependency treebank crated is not the result of automatic conversion of the LFG treebank, though the LFG treebank did serve as a basis for annotation. However, this is because they see significant mapping issues between LFG and UD (though they plan to do this automatic conversion at a later time).

13 noun classes (8 singular, 2 plural, 2 locative and 1 manner). Locus of noun class marking is on the nominal modifiers not the noun.

Determiners encode proximal and distal relations for both the speaker and addressee.

Noun classes in Wolof lack semantic coherence (citing McLaughlin, 1997).

> Wolof nouns are typically not inflected except for the genetive and possessive case

No adjective category, stative verbs used instead (similar to Swahili though there are still a small set of Adjectives in Swahili).

Towards a dependency-annotated treebank for Bambara (Aplonova & Tyers 2018)

- POS tags automatically converted (using rules), treebank handcrafted
- AS of writing, only 116 sentences with dependency annotations
- Using UD 2.0
- Bambara is predominatly isolating
- The Daba analyzer tool was used to create the original Bambara Reference corpus
- Morphological features generated by looking at both the glosses and the morphological breakdown in CBR (the reference corpus).
- compounding and derivation not treated productively so lemmas are not split into compound components
- Original reference corpus did this thing where it had multiple POS tags in cases where the POS was ambiguous. These were resolved using largely manual methods.
- All copulas were annotated as verbs? Weird choice not to have them as aux.
- Topicalization involves resumptive pronouns in Bambara

To Read

- ParGramBank
- Try to find documentation of Bambara UD treebank in paper form
- Read through information on the Yoruba UD treebank

Universal Dependencies

Universal Dependencies

A Universal Part-of-Speech Tagset (Petrov, Das, McDonald)

Abstract

- 12 universal part of speech tags
- mappings from 25 different treebank tagsets used
- Coverage of 22 different languages
- Show grammar induction for predicted part of speech tags using these "universal" tags

Introduction

- Recent interest in unsupervised POS tag induction and cross-lingual projection of POS tags.

  > Underlying these studies is the idea that a set of (coarse) syntactic POS categories exist in similar forms across languages"

- When corpora that use a standard tagset are not available, typically a mapping from fine-grained tags to a more universal POS tag set is done.
  - Das and Petrov (2011) was an example of this
- Purposes of constructing this tagset:
  - useful for evaluating unsupervised and cross-lingual taggers

    - allows for meaningful comparisons across languages when looking at supervised taggers *though the size of the corpus used can still fluctuate, at least the tagset size and distribution is roughly consistent*
    - simplifies the development of taggers across multiple languages (less annotation guideline specific information has to be utilized).
- Experiments herein:
    - POS tagging accuracy for 25 different treebanks
    - unsupervised grammar induction system for multiple languages (relying on Das and Petrov (2011) and Naseem et al, (2010).

Tagset

- Adopt a pragmatic focus, trying to find the POS categories that they expect to be most useful for users of POS taggers. - The focus is on utility for downstream tasks and grammar induction tasks
- > majority of tagsets are very fine-grained and very language specific
- Smith and Eisner (2005) made a set of 17 English POS tags from the conventional 17 *though these did not emphasize the multilingual utility of these tags*
- > McDonald and Nivre (2007) identified eight different coarse POS tags when analyzing the errors of two dependency parsers on the 13 different languages form the CoNLL shared tasks.

The tags

- NOUN
- VERB
- ADJ

- ADV
- PRON
- DET
- ADP (ADPOSITIONS)
- NUM
- CONJ
- PRT (PARTICLES)
- . (PUNCTUATION)
- X (CATCH ALL)
-
  > we did not rely on intrinsic definitions of the above categories. Instead each category is defined operationally.

  - By this, they mean that they defined these part of speech tags in their relationship to fine-grained POS tags from other treebanks
- Some tags do not occur in all languages *Adjectives don't occur in Wolof if I'm remembering that paper correctly*
  - For Korean, they treated stative verbs that would translate as adjectives in english as adjectives *this seems like a bad, Anglocentric way of doing things*.
- One important thing about these mappings is that they were established to encourage collaboraation and refinement from researcheres working on other languages (using version control etc).
- *The languages considered are very Indo-European, only 7 of the 25 treebanks are non-IE languages. However, this is probably better than most researchers were doing at the time towards including other non-IE languages*

Experiments

POS tagging accuracy comparison

-

Model: trigram markov model

  - chosen for speed, state of the art accuracy without much tuning

- Using the universal tags reduced the variance in performance across langs from 10.4 to 5.1.
- Still differences across languages
  - Japanese is very good (99% acc), Turkish worse (90.2% acc)
- > The best results are obtained by training on the original fine-graineed tags and then mapping to the UPOS tags at the end

  - > The transition model based on the universal POS tagset is less informative

Grammar induction

- Previous research on unsupervised grammar induction assumed gold POS tags. They remove this simplification using POS tags that are automatically projected from English
- Das and Petrov (2011) use cross-lingual projection to lear POS taggers without labeled data the target lang, these induced tags are used to learn unsupervised grammar.
- Using Naseem (2010)'s model where a small set of universal syntactic rules constraina bayesian model *I should read that paper if I want to make sense of what was done here*
- Using treebanks from the CoNLL-X shared task (eight indoeuropean languages used by Das and Petrov (2011))
- The method described for the grammar induction experiments in this paper are best with the gold UPOS tags performing a little better (though this wasn't the case for all languages examined, swedish for example did better with the automatically generated tags)

Universal Depedencies v1: A Multilingual Treebank Collection

(Nivre, Marneffe, Ginter, Goldberg, Hajic, Manning, McDonald, Petrov, Pyysalo, Silveira, Tsarfaty, Zeman)

Introduction

- When looking at three different, related languages (Swedish, Danish and English) that represent parallel sentences using parallel structure, the shared dependency relations have only 40% overlap
- Goals of UD:
  - > “

    > Develop cross-linguistically consistent treebank annotation for many languages

    - > capture similarities as well as idiosyncracies among typologically different langauges

    - support the following research activities:
        - comparative evaluation
        - cross-lingual learning
            - *not sure if this means human language learning or machine learning*
        - multilingual natural language processing
        - comparative linguistic studies
- This work is a fusion of several other initiatives (Stanford dependencies, Google universal dependencies, Interset morphosyntactic tag sets)

History

UD today is dependent upon prior research

- Morphological layer
    - Google universal tagset grew from cross-lingual error analysis (McDonald & Nivre 2007)
    - Interset (Zeman 2008) started as a tool for converting between the morphological tagsets of different languages
- Dependencies (syntactic layer)
    - Stanford dependencies developed for English in 2005
        - Adapted to several other languages

What other UD-like projects existed?

- Google UDT project (McDonald et al 2013) was first to combine google POS tags and Stanford dependencies
- HamleDT v2 "provided Stanford/Google annotation for 30 languages by automatically harmonizing treebanks with different native annotations"
- Universal Stanford Dependencies revised stanford dependencies for cross-linguistic use

Annotation guideline principles

- Based upon dependencies
- based upon lexicalism
    - > words are the basic units of grammatical annotation
- syntactic wordhood != orthographic wordhood
- Recoverability principle
    - > there should be a transparent relation between the original textual representation and the linguistically motivated word segmentation
- maximize the parallelism between languages
    - ensuring the same construction is annotated in teh same way across langauges
    - don't want to annotate thigns that do not exist in a language simply because that's how they work in other languages *this seems to conflict with the annotation of Korean*

stative verbs as Adj for the Universal POS tagset paper

- > use a universal pool of structural and functional categories that languages select from

- > possible to refine the analysis by adding language-specific subtypes

Word segmentation

- Clitics split off
- contractions are undone *seems like a strange decision. why not split up compounds too if you're undoing contractions*
- > UD currently does not allow words with spaces

Morphology

Lemma

No guidelines provided for what the lemmas should look like. E.g. should lemmas include derivational morphemes, what should you do for suppletives etc.

Part of speech tag

- 17 part of speech tags, a fixed set for all languages to draw from but not all tags need to be present in all languages

Morphological features

- Based on the interset ssytem

- Each feature is associated with a set of possible values

Syntax

- 40 different grammatical relations for version 1.0
- 3 types of structure:
  - nominals
  - clauses
  - modifier words
- Distinction between core arguments and other dependents which is different from complements vs adjuncts.
  - Core arguments are subjects and objects, *other arguments are non-core even if they are required by the verb*
- The attachement point of a relation is crucial
  - For example, an adverbial clause that modifies a noun is `acl`, an adverbial clause that modifies a predicate is `advcl`
- Rich collection of noun dependents
- Relations for non-edited/informal text also included
  - e.g. reparandum
  - goeswith
- compounding
  - mwe for fixed expressions containing function words *largely corresponds to `fixed` in UD v2*
  - name for names consisting of multiple propoer nouns *largely corresponds to `flat` in UD v2*
  - compound is used for any kind of lexical compounding *still `compound` in UD v2*

mwe and name are both left headed with a flat structure (e.g. all are connected to the left-most part of the name or mwe). *This is carried over to `fixed` and `flat` in UD v2 which means I need to fix some of my names that I've annotated*

Relations between content words

- Priority is given for dependency relations between content words
    - Increases chances of parallel structure between languages because functional words can just be indicated using morphology or other non-syntactic means

    > The UD view is that we need to recognize both lexical and functional heads, but in order to maximize parallelism across languages, only lexical heads are inferable from the topology of our tree structures

- Very close to the view of Tesniere (1959) *the OG dependency grammar*

Language-specific relations

- UD allows the use of language-specific relations to capture extra stuff

CG to Dependency Parse

Reusing Grammatical Resources for New Languages

Lene Antonsen, Trond Trosterud, Linda Wiechetek

Takeaway

Machine-readable grammars can be more easily applied to new langauges if they are working with higher levels of analysis. Working with morphophonology, the grammatical differences between languages preclude the reuse of analyses.

> We argue that portability here takes the form of reusing smaller modules of the grammar

Hopefully the paper expands on that because that statement doesn't make any sense

Languages

- North Lule and South Sami
  - Uralic language
  - Not very agglutinative
- Faroese
  - Germanic language
  - Four case system
- Greenlandic
  - Eskimo-Aleut language
  - Polysynthetic

Technical background

- Using existing resources developed by the University of Tromso.
  - Morphological analyzers
  - Constraint Grammar parsers

Reusing grammar

- Blick (2006) argues for using bootstrapping techniques to reuse grammar instead of appealing to statistical systems. *This fell by the wayside, everyone uses statistical methods now*

The bottom of the analysis

- The level of analysis that is close to the language substance cannot be directly used
- > “

> Even though different languages do not have the eact same morphological processes, they may have the same process *types*

- Rules are written in a modular fashion so they can easily be adapted to new languages
    - For example, consonant gradiation processes are very common, the particulars of the rule may need to change but the module design helps guide the changes that need to be made.

Disambiguation

Mapping of syntactic tags

- Large number of tags needed due to the free word order of Sami languages
    - For example, four different subject tags needed specifying whether the verb is finite, whether elipsis of verb has occured, whether the finite verb is to the left or to the right etc.
-

The top of the analysis

This is the part that's relevant to me

- Using a constraint grammar module
- > Syntactic tags for verbs are substituted by other tags (according to clause-type) in order to make it easier to annotate dependency across clauses
- Descibes difficulties finding the "head" of the sentence (think they mean root), when dealing with ellipses. This is definitely an issue as well in UD

> Still the analyzer retains very good accuracy for the dependency analysis: 0.99

- This is for Sami
- Table 5 say this is actually f-score?
- How is this scored? Are they scoring the flat descriptors in the visl format (e.g. #5->0)
- Use pairs of substitution and setparent rules

Bootstrapping

- Go through small modifications to the rules to consider Faroese specific phenomenon.
- Show the specific increases in performance with each new difference that is considered (e.g. when substituting the Relative pronouns that begin subordinate clauses in Sami with the CS that begins relative clauses in Faroese, the accuracy goes up to 96)

Estonian Dependency Treebank: from Constraint Grammar Tagset to Universal Dependencies

Kadri Muischnek, Kaili Müürisep, Tiina Puolakainen

Central topic

Dependency treebank in Universal Dependencies formalism adapted from an existing dependency treebank for Estonian. this adaptation was doen semi-automatically using a Constraint Grammar transfer rule system.

Methodology

Structure of annotations

The Estonian Dependency Treebank (DT) is annotated in Constraint Grammar style. There are three layers:

- morphological
- surface syntactic
- dependency

This is an example word tag in a larger sentence (for more information see Figure 1 in the paper).

```
"<lamnast>"
   "lamnas" Lt S com sg part @<Q #6->5
```

> “The used set of syntactic relations derives from Constraint Grammar, but the definitions of syntactic relations...are based on an academic description of Estonian grammar

Differences between UD and EDT annotation

Both EDT and UD adopt dependency grammar-based annotation guidelines. However, different syntactic relations are used and some phenomena are analyzed differently.

POS tags

No DET tag in estonian UDT, smilar decision made for The Fininnish UDT. PART not used because these things are currently tagged as adverbs or pronouns and it would require manual effort to retag them.

No discussion of annotation of morphological features.

Ditransitives are not used as there are no grammatical descriptions of Estonian that describes ditransitives in Estonian.

EDT distinguishes between finite and non-finite (*subordinate*) clauses with finite clauses not indicating the syntactic relation between the head of the finite clause and the main clause *what are they doing here then? This is very unclear in this paper, maybe I need to read the paper for the EDT in order to make sense of this*.

EDT annotated modals and other auxiliaries as multi-word predicates. Many of these are set up as complementary clauses with ccomp and xcomp in UD instead.

Primacy of content words in UD causes a large number of changes. EDT did a lot of relations between functional words. For example, nouns in a prepositional phrase were dependents of the preposition, while the preposition was dependent on the larger context. In UD, this has to be changed because dependency relations need to be between content words.

Conversion procedure

- Rearrange subtrees, find connections between UD and EDT *I thought this was manual exploration of differences first, but it does appear this is the actual tree rewriting*
    - Using Vislcg3 *like I intend to*
- Convert from CG3 format (default in ED%) to CONLL-U, convert pos tags, morphological features using simple mapping.
- Formal checks to verify there is one and only one root, verify valid dat, all fields filled in.

Findings

Estimation of conversion quality:

- Used MaltEval
- UAS of 96.3
- LAS of 98.4

- annotation of punctuation marks was an issue.
- ccomp is the most error prone dependency relation at 64.3%

UD's emphasis on dependencies between content words results in projectivity (often). Where EDT was non-projective, the UD version is projective.

Follow up readings

Bantu NLP

Learning Morphosyntactic analyzers from the bible via iterative annotation projection across 26 languages

Garrett Nicolai and David Yarowsky

Central topic

Morphlogical analysis and lemmatization using English taggers, cross-linguistic projection and then an iterative discovery, constraint, and training process.

Methodology

Findings

Follow up readings