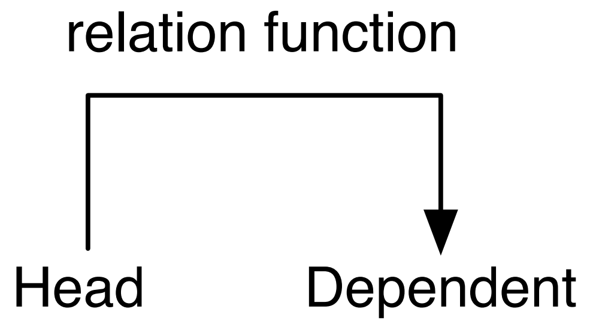
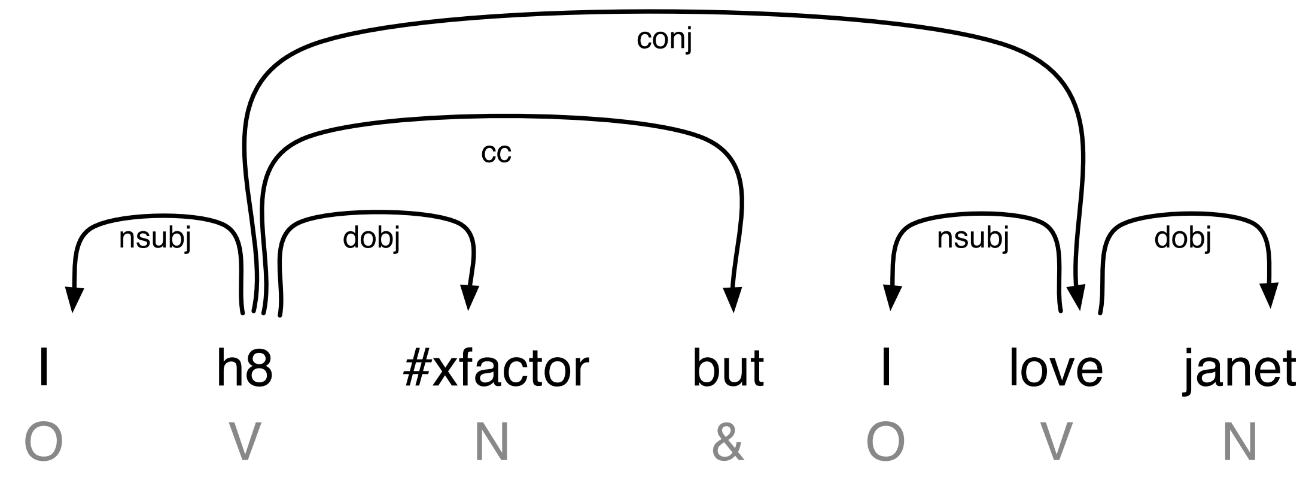
Extracting syntactic structure from microblogs

by adapting dependency parsing to Twitter

What is dependency parsing?



- A dependency tree consists of binary asymmetric relations between words, showing that a word depends in some way on another.
- It uses these relations to show how the words in a sentence interact with one another.
- The parser requires that the sentence is first split into a list of words, each with their Part of Speech (POS) tags assigned (e.g. “V” for “verb”).



nsubj	nominal subject
dobj	direct object
cc	coord. conjunction
conj	conjunct

O	pronoun
V	verb
N	noun
&	coord. conjunction

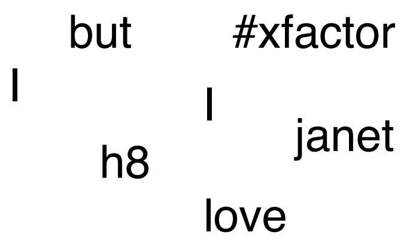
Why is it useful?

Deeper analysis

Information extraction

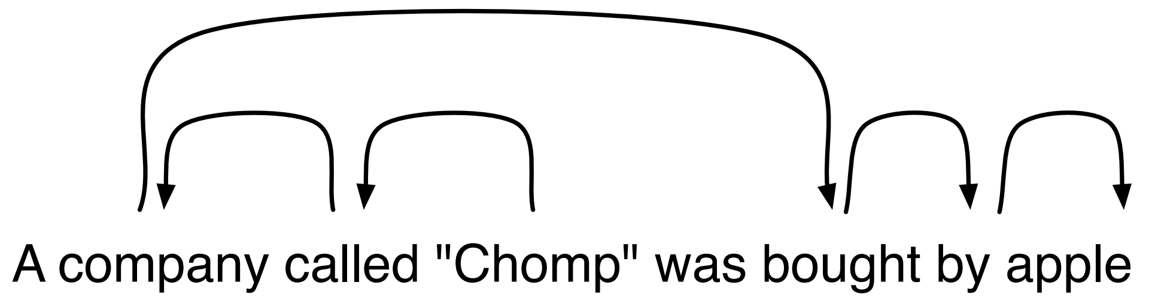
Machine translation

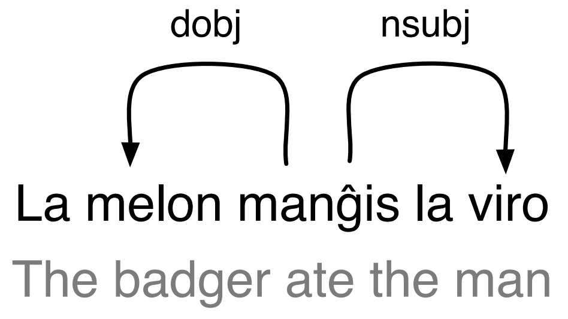
Semantic role labelling



A classifier using the “bag-of-words” approach, cannot determine how the words relate to each other. How do we know what the hate/love applies to?

We want to assert the predicate: “buy(Apple, Chomp)”, this is much easier when the relations between words are known.





The word-for-word translation in grey, is shown to be wrong by the dependency relations. It actually says “The man ate the badger”.

The challenge

Aint watching xfactor for shit qithout @AmeliaLilyOffic on it no more #fix

RT @komz_x #Xfactor time... Can't wait to see Marcus Collins and Sophie Habibis :)

OMG Xfactor, #GOFRANKIE.

#XFactor #VOTEJANET #WELOVEJANET. Vote janet

I can't to see @FrankieCocozza

Eye Love Nu Vibe. #Xfactor

Nu vibe better of worked hard this wk cos they were shit last wk #xfactor

NU VIBE NEED TO GO KILLING THAT SONG #XFactor

Nuvibe. No just no.

Nu Vibe are murdering Ross and Rachels song.

Mmmm #nuvibe arnt on their a game!

In the above, spot examples of:

Sentence fragments

Incorrect spelling

Missing punctuation

Missing words

Twitter-specific notation

Emoticons

Multiple words concatenated inside a Twitter hashtag

Slang

Colloquial acronyms

Grammar errors

Abbreviations

Unusual usage of words

Unusual sentence structure

Hashtags used as part of sentence instead of topic classification

Other major roadblocks:

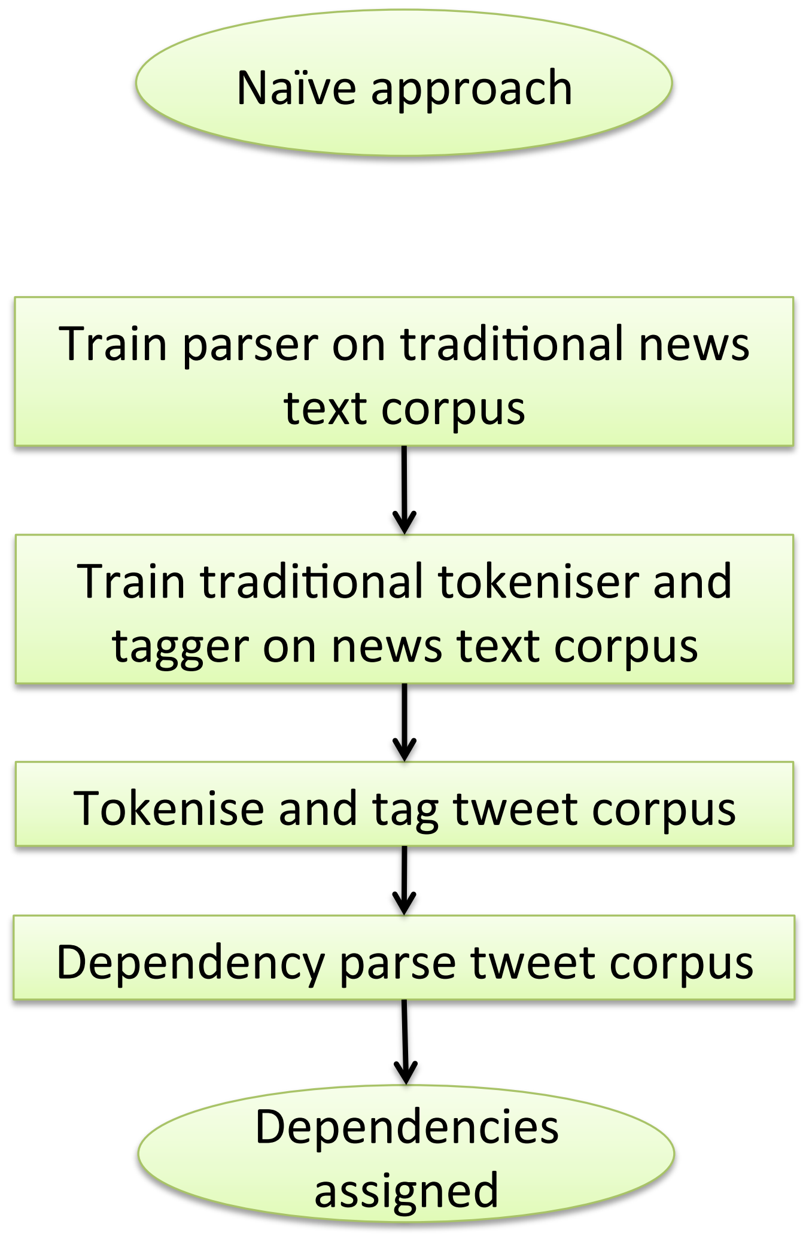
For the parser to work, it must see many training examples of sentences annotated with dependency trees, but the only available data is news text.

Must tokenise tweets, but traditional tokenisers aren’t used to seeing @-tags, hashtags, emoticons, etc.

Must POS-tag tweets, but traditional taggers are also trained on news text

If we use a non-traditional more Twitter-specific POS tagger and tokeniser, it will produce sentences and features that are completely different from those that the parser was trained on.

The approach



Naïve approach

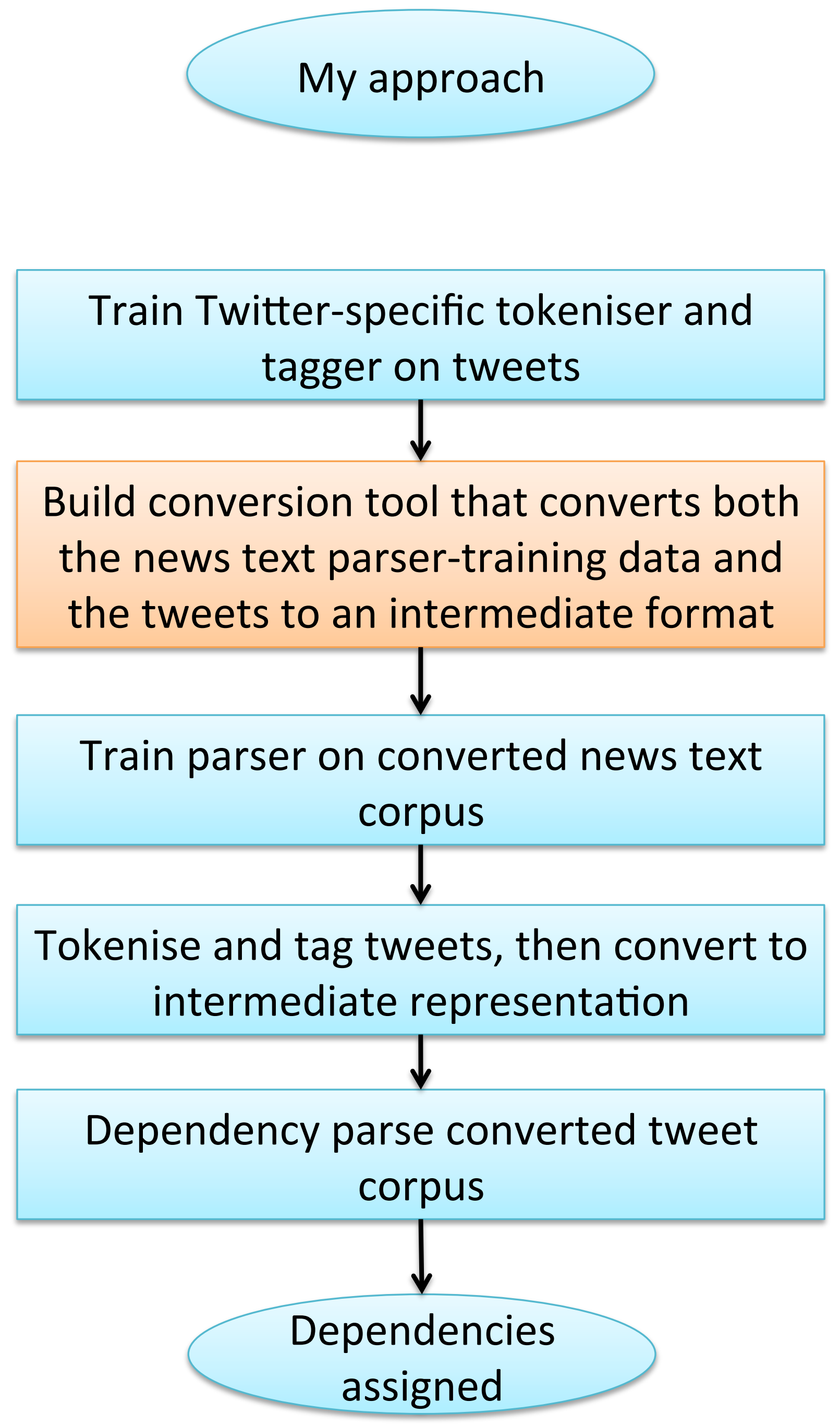
Train parser on traditional news text corpus

Train traditional tokeniser and tagger on news text corpus

Tokenise and tag tweet corpus

Dependency parse tweet corpus

Dependencies assigned



My approach

Train Twitter-specific tokeniser and tagger on tweets

Build conversion tool that converts both the news text parser-training data and the tweets to an intermediate format

Train parser on converted news text corpus

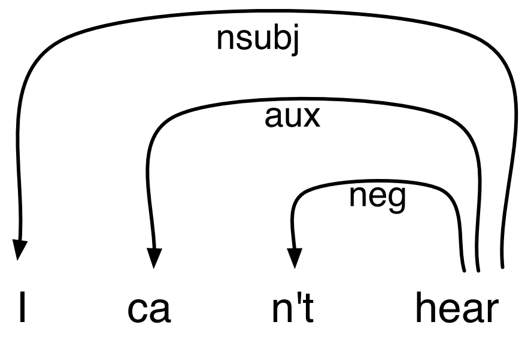
Tokenise and tag tweets, then convert to intermediate representation

Dependency parse converted tweet corpus

Dependencies assigned

The conversion tool’s responsibilities

Split into several tokens contractions like “can’t” (with or without the apostrophe)



Expand abbreviations like “iono” to “I do not know”

Remove twitter notation that isn’t part of the sentence, like “retweet” data and URLs

Remove only those hashtags that are just assigning a topic to the entire tweet

Convert Twitter-specific POS-tags to their nearest equivalent in what would have appeared in the training news text. E.g. @-tags like “@janet_devlin” would be tagged as proper nouns.

Andrew D. Robertson
Text Analytics Group (TAG)