Supplementary Notes: Model Architectures for Quotation Detection

Christian Scheible, Roman Klinger and **Sebastian Padó**
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
{scheibcn,klinger,pado}@ims.uni-stuttgart.de

1 Feature Templates

As we want our model to be easily reproducible, we include formal specifications of our feature templates in this appendix. We use three different sets of templates: one for cue identification, one for span identification, and a third for global span features. We define all templates in detail in the sections below.

Binning Our stacked binnings use the following intervals and stack both upwards and downwards: 1, 2, 3, 4, 5, 6, 7, 8, 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, 100. Interval binnings use the following interval boundaries: 0, 5, 10, 20, 40, 60, 80, 100.

Shape feature The shape of a token is determined by replacing each character by a placeholder:

- `X` for uppercase characters
- `x` for lowercase characters
- `0` for digits
- The character itself for all others

We then collapse sequences of equal characters longer than 3 tokens (e.g., `Xxxxxx` becomes `Xxxx`). This procedure is adapted from the shape feature from FACTORIE.[1]

1.1 Cue Features

For *cue identification*, we use the following templates for each token t_i at position i, which are mostly derived from Pareti (2015):

C1. Surface form, lemma, and PoS tag for all tokens t_{i+n} for $n \in -5, \ldots, +5$. This feature is conjoined with an indicator for whether t_{i+n} is in the same sentence as t_i.
C2. Bigrams of surface form, lemma, and PoS tag with t_{i-1} and t_{i+1}
C3. Shape of t_i
C4. Are t_{i-1} or t_{i+1} capitalized?
C5. Is any of t_{i+n}, $n \in -5, \ldots, +5$, named entity?
C6. Does a quotation mark open or close at t_i (determined by counting)? Is t_i within quotation marks?
C7. Is t_i in the list of reporting verbs by Krestel et al. (2008)?
C8. Is t_i in the list of noun cue verbs by Pareti (2015)?
C9. Is t_i in the list of titles (will be published with paper)?
C10. Is t_i in the list of WordNet persons (will be published with paper)?
C11. Is t_i in the list of WordNet organizations (will be published with paper)?
C12. All VerbNet classes of which t_i is a member
C13. Do a sentence, paragraph, or the document begin or end at t_i, t_{i-1}, or t_{i+1}?
C14. Length of the sentence containing t_i
C15. Does the sentence contain t_i a pronoun/named entity/quotation mark?
C16. Distance to sentence begin and end (count and interval bins)
C17. Does a syntactic constituent starts or ends at t_i?
C18. Level of t_i in the constituent tree
C19. Label of the highest constituent in the tree starting at t_i
C20. Level of the highest constituent in the tree starting at t_i
C21. Conjunction of C19&C20
C22. Label of the parent constituent of t_i (unindexed as well as indexed with its level)
C23. Dependency relation with parent of t_i (with and without parent surface form)
C24. Dependency relation with any child of t_i (with and without child surface form)
C25. Any conjunction of C6, C15, C17

[1] `https://github.com/factorie/factorie/blob/master/src/main/scala/cc/factorie/app/strings/package.scala`

1.2 Token-Level Span Features

The *boundary identification* and *content span* models have access all the features of the cue classifier as well as additional feature templates that make use of the output of the cue classifier:

S1. Is a direct or indirect dependency parent of t_i in the reporting verb list?
S2. Is a direct or indirect dependency parent of t_i in the noun cue list?
S3. Is a direct or indirect dependency parent of t_i the phrase "according to"?
S4. Was a direct or indirect dependency parent of t_i classified cue?
S5. Was a direct dependency parent of t_i classified cue?
S6. Was any t_{i+n} for $n \in -5, \ldots, +5$ classified cue? If so, which?
S7. Distance to the previous and next cue up to 50 tokens (stacked bins in both directions as well as interval bins)
S8. Does the sentence containing t_i have a cue?
S9. Conjunction of 8 and 25

1.3 Global Span Features

The semi-Markov model can make use of span-global features which we cannot include in the token-level models. These templates are shown below for a span between the tokens t_b and t_e.

G1. Numbers of named entities, lowercased tokens, commas, and pronouns among the tokens t_{b+1} through t_{n-1}
G2. Binned percentage of tokens that depend on a cue
G3. Do t_b and t_e both depend on a cue?
G4. Location of the closest cue (left/right?)
G5. Percentage of tokens on the span that are dependents of the closest cue
G6. Number of tokens classified as cue between t_b and t_e (cue overlap)
G7. Does a cue occur before t_b (within the same sentence)?
G8. Does a cue occur after t_e (within the same sentence)?
G9. Conjunction of G7 and G8
G10. Do both the first and the last token depend on a cue?
G11. Length of the span: interval bin as well as absolute number if shorter than or equal to 5 tokens
G12. Number of sentences covered by the span
G13. Does the span matches a sentence exactly?
G14. Is t_b the beginning of a sentence and is t_{e+1} the end of that sentence?
G15. Does the span match a single constituent exactly?
G16. Does the span match multiple constituents exactly?
G17. Is the span direct, indirect, or mixed?
G18. Is the # of quotation marks in the span odd or even?
G19. Is the span is direct and does it contain more than two quotation marks?

2 Features using Lists

Following Pareti (2015), our feature sets make use of lists persons, organizations, titles, verb and noun cues, as well as verb classes. As we were unable to obtain the original lists, we created our own resources. For persons and organizations, we collected the transitive closure of hyponyms of the words *person* and *organization*, respectively, from WordNet. We manually compiled a list of titles from Wikipedia.[2] Verb cues were taken directly from Krestel et al. (2008). Noun cues are listed in Appendix C of Pareti (2015). We further extracted a mapping of verbs to verb classes from VerbNet. These resources are available in electronic form at `http://www.ims.uni-stuttgart.de/data/qsample`.

References

Ralf Krestel, Sabine Bergler, and René Witte. 2008. Minding the source: Automatic tagging of reported speech in newspaper articles. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 2823–2828, Marrakech, Morocco.

Silvia Pareti. 2015. *Attribution: A Computational Approach.* Ph.D. thesis, University of Edinburgh.

[2] `https://en.wikipedia.org/wiki/Title`