Effective Use of Context in Noisy Entity Linking

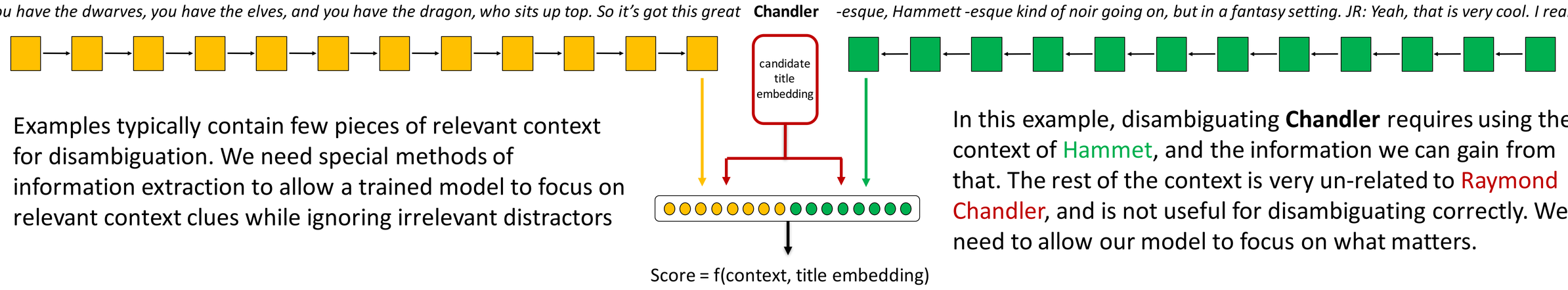
David Mueller
Johns Hopkins University
dam@jhu.edu

Greg Durrett
University of Texas at Austin
gdurrett@cs.utexas.edu

Introduction

To disambiguate between concepts, entity linking systems need to **distill cues from a mention's textual context**. We investigate several techniques for using these cues in the context of noisy entity linking on short texts. We start with a state-of-the-art attention-based model from prior work. This prior model **fails to identify some of the most indicative context words**, especially those exhibiting lexical overlap with the true title. We investigate extensions of this model: **using convolutional networks over characters still leaves it largely unable to pick up on these cues compared to sparse features that target them directly**, indicating that automatically learning how to identify relevant character-level context features is a hard problem.

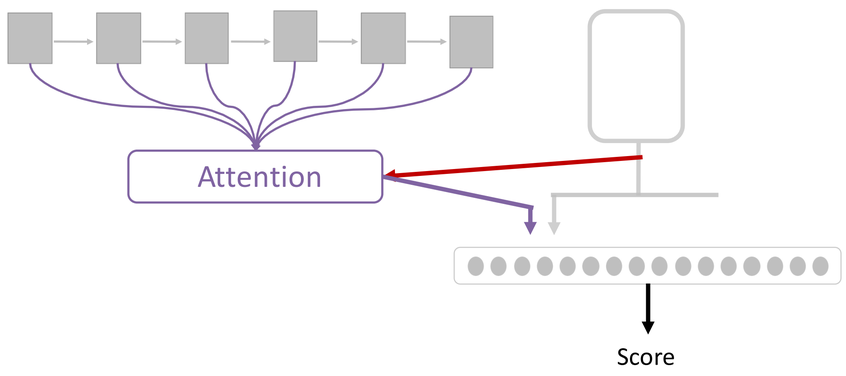
Base Model



Examples typically contain few pieces of relevant context for disambiguation. We need special methods of information extraction to allow a trained model to focus on relevant context clues while ignoring irrelevant distractors

In this example, disambiguating **Chandler** requires using the context of Hammet, and the information we can gain from that. The rest of the context is very un-related to Raymond Chandler, and is not useful for disambiguating correctly. We need to allow our model to focus on what matters.

Attention



Chess Tactics by John Bain to help tactics . Logical Chess: Move By Move by Irving Chernev to help with

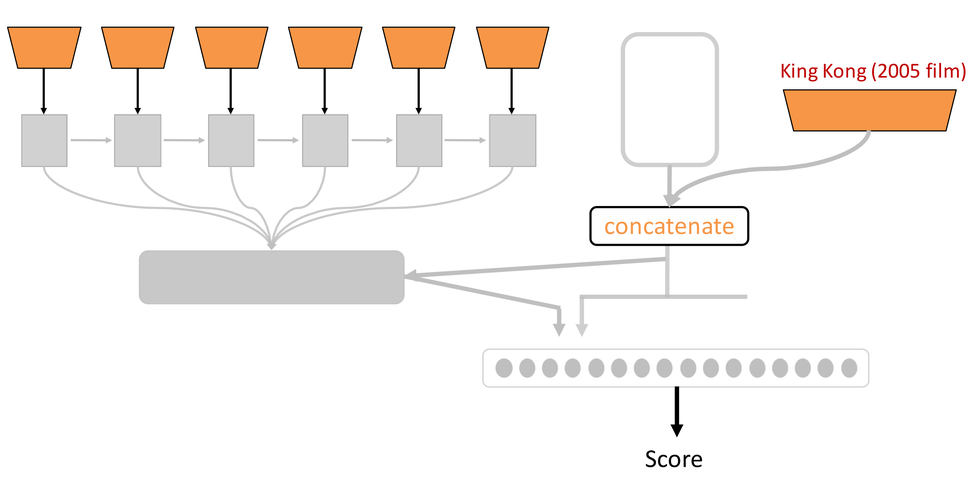
strategy

Posted 8 months ago Comments 7 notes chess , hobbies , interests , View the discussion thread Blog

Attention should help the model focus on important context clues while putting less weight on distracting terms and words.

We want our model to focus on words related to **Chess strategy**, while ignoring noisy web-page context that often has random terms in it.

Character-level CNNs



Cooper, Lordi and Sonic Altar with special effects provided by WETA Workshop of Lord of the Rings and

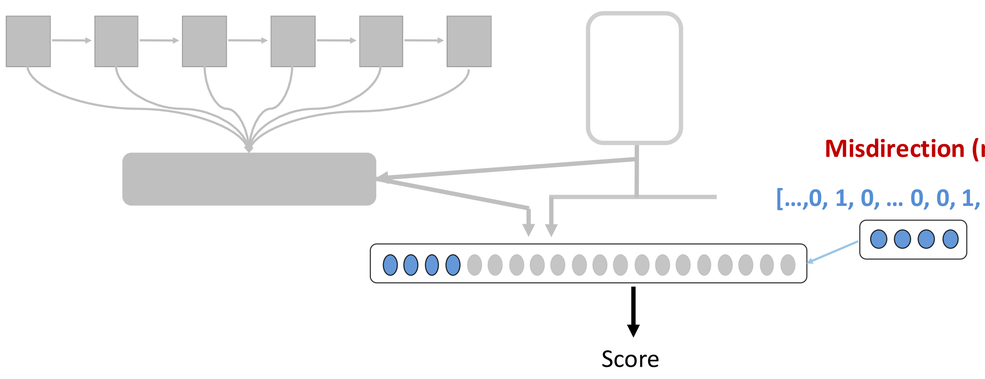
King Kong

fame. Throughout the summer of 2008, Kiss headlined festivals as well as their own shows and played

Some examples seem to require context clues coming from a lexical level, due to poor token-level embeddings caused by misspelled or rare tokens.

We should prefer King Kong (2005 film) over King Kong (1933 film) due to the context referencing 2008, but our model doesn't have a good semantic representation of the token 2008.

Lexical Features



post Labels: Health care reform, health care reform plan, Misdirection to conceal . . . "the inevitable"

Pos:6-11::Match:Exact::Context:Left

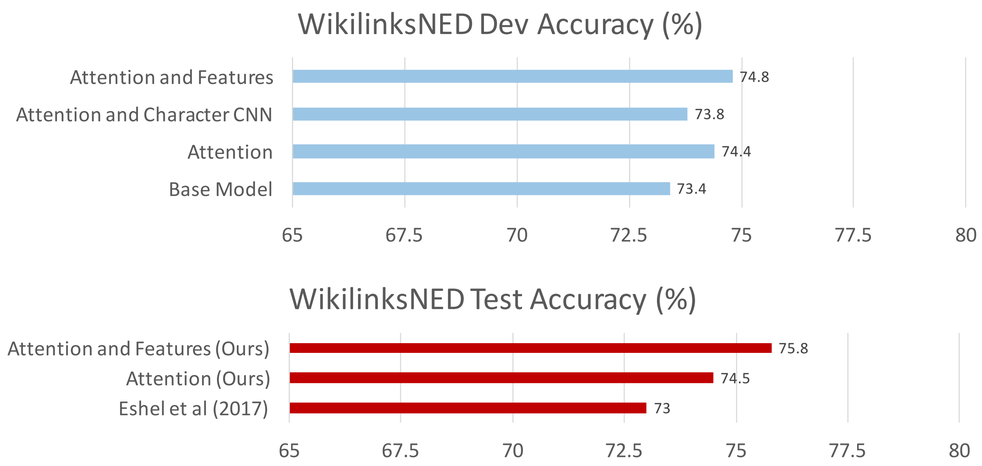
Misdirection

Pos:5::Match:Substr::Context:Right

is a technique used by magicians to divert the attention of the spectator from where the trick is really.

We can also feed lexical overlap between the Wikipedia article title and context directly into the model using coarse features, if the context is too noisy for the model to capture it on its own. Here, the attention fails to attend to *magician* as the most relevant context clue when evaluating Misdirection(magic) and gets the example wrong. Features identify this overlap directly.

Results



- We observed that, while attention helps in this setting, it does not always pick up on the correct context clues**, even when those clues exhibit very obvious surface overlap with the correct entity title**.
- These models can perform better when augmented with **sparse features explicitly targeting this kind of lexical overlap**: our system using these features achieves state-of-the-art disambiguation accuracy on the WikilinksNED dataset.
- By contrast, **automatically learning fine-grained character-level features with CNNs in this context is hard**. We found that the, while the CNNs did slightly aid attention in focusing on words exhibiting the lexical features described above, it wasn't significant enough to aid in performance. (See paper for more details).