A Simple and Effective biLSTM Approach to Aspect-Based Sentiment Analysis in Social Media Customer Feedback

Simon Clematide

The Tasks C and D of GermEval 2017 ABSA

Domain German-language customer feedback about "Deutsche Bahn (DB)"

Task C Predict aspects (and their sentiment) on the **document** level!

Task D Predict aspects (and their sentiment) on the **mention/token** level!

Annotation Example

Stand-off annotation XML format

- NULL target for document-level aspects
- Aspect-level and document-level sentiment classes (positive, neutral, negative)

```
<Document ...><Opinions>
<Opinion category="Allgemein" from="0" to="0"
 target="NULL" polarity="negative"/>
<Opinion category="Sonstige_Unregelmässigkeiten" from="5" to="20"
 target="Weichen Störung" polarity="negative"/>
</Opinions>
<relevance>true</relevance><sentiment>negative</sentiment>
<text>Juhu Weichen Störung!  Ich liebe die Bahn ...  Nicht
-.-</text>
</Document>
```

Annotations as Sequence Labels

Aspect classification as sequence labeling of lowercased tokens:

```
juhu/O weichen/Sonstige_Unregelmässigkeiten:negative
störung/Sonstige_Unregelmässigkeiten:negative !/O
ich/O liebe/O die/O bahn/O .../O nicht/O -.-/O
__D__/Allgemein:negative
```

- Dummy token __D__ represents **document-level aspect/sentiment**.
- Documents can have more than one label in the original data: **multi-label multi-class problem**
- We simplify to single-label problem by reducing multi-label cases to the **most frequent label** in training data.
- Annotations of **subwords** are projected to containing token
- No IOB encoding, just **plain IO** for multi-token annotations

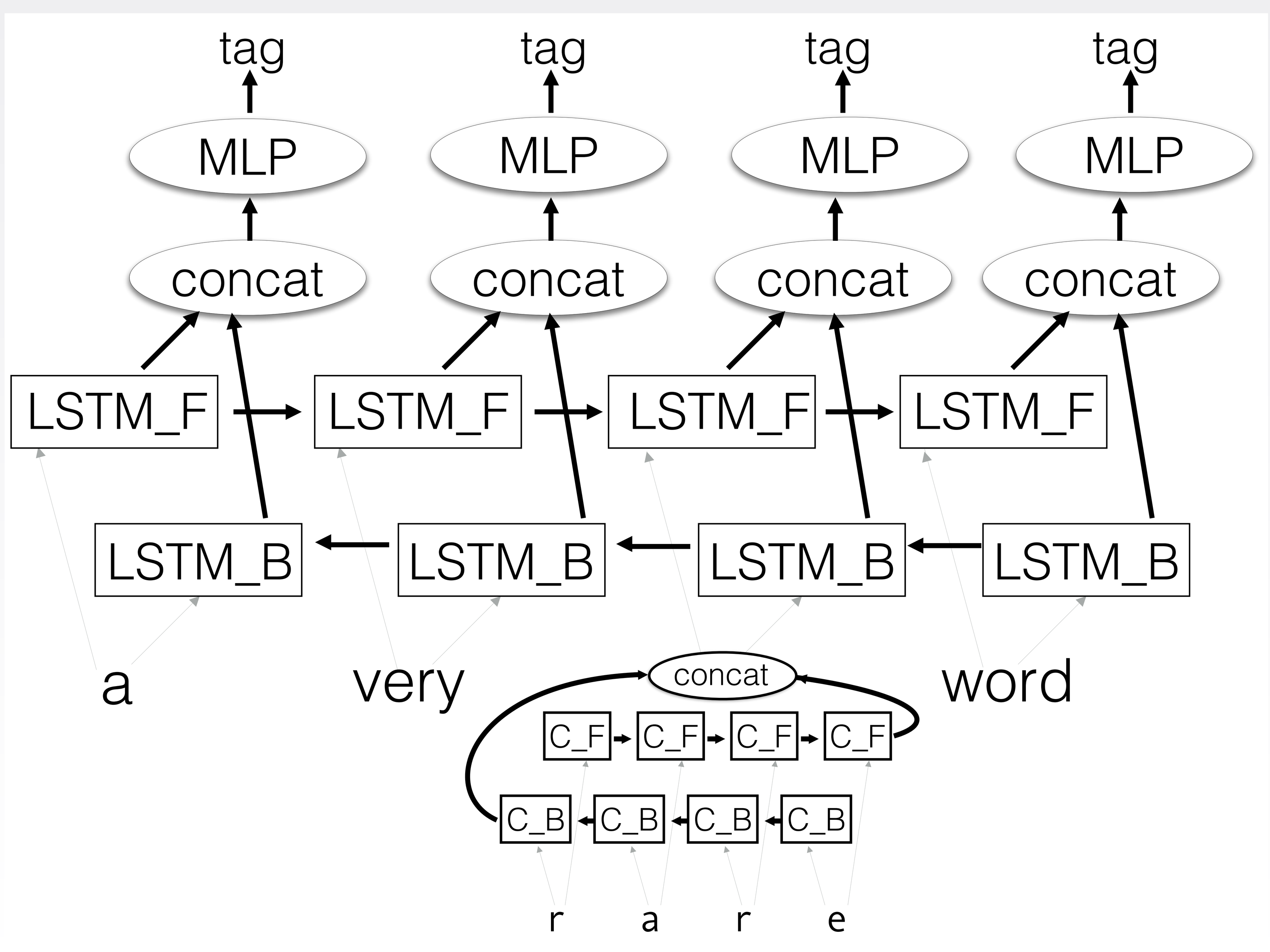
Important Data Set Properties

Imbalanced and sparse classes

- 17,758 **neutral**, 6,911 **negative**, 1,540 **positive** documents
- **20 aspect classes**: out of 21,772 aspect annotations 68.5% are GENERAL; the top 10 real aspect categories cover 29.2%, the long tail only 2.32%
- Majority of tokens belongs to the **uninteresting class O**.
- Combination of aspects and sentiments labels A:S on training data results in **59 classes**.

General Neural Architecture

Dynamically switch between token and character embeddings:



Our Simple BiLSTM Method

Forward LSTM encoding of a task-specifically embedded input token sequence $(\mathbf{x}_1, \ldots, \mathbf{x}_T)$:

$$F = (\mathbf{h}_1, \ldots, \mathbf{h}_T) = LSTM((\mathbf{x}_1, \ldots, \mathbf{x}_T))$$

And a **backward LSTM** encoding $B = LSTM((\mathbf{x}_T, \ldots, \mathbf{x}_1))$ concatenates elementwise into a biLSTM representation:

$$(\mathbf{b}_1, \ldots, \mathbf{b}_T) = biLSTM((\mathbf{x}_1, \ldots, \mathbf{x}_T)) = ([F_1; B_1^{-1}], \ldots, [F_T; B_T^{-1}])$$

(B^{-1} means reversed B)

Contextualized biLSTM representation $\mathbf{b}_i$ of input word $\mathbf{x}_i$ goes into multilayer perceptron (MLP) and softmax probabilistic classification layer:

$$P(y_i^k) = \text{softmax}^k(MLP(\mathbf{b}_i))$$

Rare words (≤ 2) have character-level biLSTM representation:

$$biLSTM_{char}((\mathbf{x}_1, \ldots, \mathbf{x}_T)) = [B_T; F_T]$$

ADAM training **without mini-batching** using cross entropy loss and early stopping based on F-Score for non-O class labels.
Aspect-favoring **voting ensemble** using 24 models: if at least 33% of the models predict a non-O label, take it.

Results Task C

Best Shared Task Submissions and LT-ABSA system by organizers compared by F-Score on synchronic (SYN) and diachronic (DIA) test set.

Task C	SYN		DIA	
System	A	A:S	A	A:S
Majority bsl.	44.3	31.5	45.6	38.4
Organizers' bsl.	48.1	32.2	49.5	38.9
Mishra	42.1	34.9	46	40.1
Lee (best run)	48.2	35.4	n/a	n/a
LT-ABSA	**53.7**	39.6	**55.6**	42.4
Our A	49.0		53.2	
Our A:S	49.6	**39.8**	53.6	**44.7**

A=aspect, S=sentiment, bsl.=baseline

Comments

- We predict set of aspects per document (evaluation script expects bag of aspects).
- Training on **combined label A:S** is beneficial
- New state of the art for A:S predictions

Results Task D

Task D (=OTE)	SYN		DIA	
System	exact	overl.	exact	overl.
Organizers' bsl.	17.0	23.7	21.6	27.1
Mishra	22.0	22.1	28.1	28.2
Lee (best run)	20.3	34.8	n/a	n/a
LT-ABSA	22.9	30.6	30.1	36.5
Our	**36.8**	**37.5**	**44.4**	**45.2**

Improves state-of-the-art results by **2.7-14.3 points**.

Key Findings

- Use simple IO encoding of document and token level aspect classes
- Use task-specific word and character embeddings for a very domain-specific task (given the generous amount of training data)
- Jointly training for aspect and sentiment is beneficial
- Mix word and character-level representations
- Use (reasonably) generous ensembling with a recall-oriented voting threshold for aspects (given the imbalanced class distribution)

Code and Acknowledgments

Our code: `https://github.com/simon-clematide/konvens-2018-german-absa`
Derived from DyNet POS tagger code: `https://github.com/clab/dynet_tutorial_examples/blob/master/tutorial_bilstm_tagger.py`
Thanks to the GermEval 2017 organizers for creating these data sets and tasks!
Presented at Konvens 2018, Vienna

Institute of Computational Linguistics

University of Zurich[UZH]
http://www.cl.uzh.ch
Andreastrasse 15, CH-8050 Zurich
Contact: simon.clematide@cl.uzh.ch