iTalk

Chris Lin clin17@stanford.edu
Qian (Sarah) Mu sarahmu@stanford.edu
Yi Shao yishao@stanford.edu

Summary

In text-to-speech (TTS) synthesis, the written form of a text is transformed to its spoken form. In our project, we developed a 3-component transformation pipeline. In the pipeline, text class such as “CARDINAL” for text “2007” is useful for applying the appropriate grammar rule. Test accuracy reached 98.88% with token-level Naïve Bayes (NB) combined with Support Vector Machine (SVM) classifier and grammar rules.

Data

Kaggle.com provides 670,000 sentences with around 9 million written tokens and corresponding ground-truth token class labels and spoken forms [1-2]. The token classes are listed below:

CARDINAL	TELEPHONE	PLAIN	LETTERS
DATE	MEASURE	PUNCT	ADDRESS
DECIMAL	MONEY	ELECTRONIC	TIME
DIGIT	ORDINAL	FRACTION	VERBATIM

Features

To build classifiers for token class, we need to represent the tokens in numeric values. We used the bag-of-words model with a bag of 162 English characters found in our data set (discarding 2,933 non-English characters). We constructed term frequency (TF) and L2-normalized term frequency-inverse document frequency (TF-IDF). These features are appropriate as the token classes would have different distributions of characters (e.g. the “DIGIT” class has more numbers than “PLAIN”).

Models

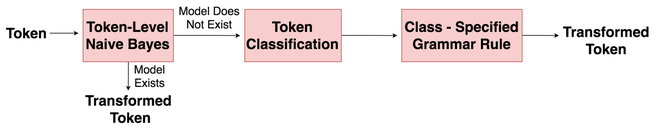
I. Direct Transformation – **NB:**
Construct a set of token-to-token NB models. Predict as original written form if unseen before.

$$p(x_1, ..., x_n|y) = \prod_{i=1}^{n} p(x_i|y)$$

II. **L2-SVM** with TF-IDF (one-vs-all scheme):
Construct multi-class SVM model for TF-IDF-to-class prediction:

$$\begin{aligned} &minimize \quad \frac{1}{2}||w||^2 + \frac{C}{2}\sum_{i=1}^{M}\xi_i^2 \\ &subject\ to \quad y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i \\ &\xi_i \geq 0 \end{aligned}$$

3 Component System



Results

Model	Training	Dev.	Test
Set size	6,600,000	1,000,000	1,300,000
Token-level NB[1]	99.81%	98.97%	98.88% (Token-level NB + SVM classifier)
Token-level NB + SVM classifier[2]	99.81%	99.36%	
Token-level NB + NB classifier[3]	99.81%	99.32%	

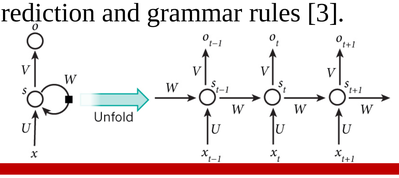
Note:
1.Benchmark accuracy: 93.34% (spoken=written)
2.Used TF-IDF as input and 3 component system
3.Used TF as input and 3 component system

Discussion

1. Token-level NB was able to predict spoken form with high accuracy. However, NB could not predict new written forms. Thus, token-level NB was combined with SVM classifier and grammar rules.
2. In token-level SVM classifier, the majority of error came from the inability to classify numbers.
3. Tuning parameters for SVM showed that overfitting training set was avoided, and unbalanced model with penalty parameter of 0.5 was the best.
4. In our error analysis, perfecting token-level NB improved accuracy by 0.20%; perfecting SVM classifier improved accuracy by 0.18%; perfecting grammar rules improved accuracy by 0.26%.

Future

Based on our error analysis, grammar rules need to be improved. We plan to try RNN (recurrent neural network) for both class prediction and grammar rules [3].



References

[1] Data downloaded from https://www.kaggle.com. The Text Normalization Challenge - English Language sponsored by Google's Text Normalization Research Group.
[2] P. Ebden and R. Sproat, “The Kestrel TTS text normalization system,” Nat. Lang. Eng., 2014.
[3] R. Sproat and N. Jaitly, “RNN Approaches to Text Normalization : A Challenge,” CoRR, vol. 1611.00068, 2016.