Toward A Better Image Captioning Metric & Kanji-based word representations

Clippers 2/17/2016

Evan Jaffe

Goals: Brainstorming, next steps

Background on automatic image captioning and its evaluation

Motivation for alternate metric

My metric, preliminary results

Current challenges

Goals: Brainstorming, next steps

Background on Japanese morphology

Possible tokenizations

Motivation for kanji-based tokenization

Possible evaluations

Automatic Image Captioning

Task: Generate a textual description of an image

Some recent models: RNNs for visual saliency and attention (Xu et al. 2015), projection of visual and language features in/out of a shared space (Socher et al. 2013; Hodosh, Young and Hockenmaier, 2013;Vinyals et al. 2015; Karpathy and Fei-Fei 2015, inter alia)

Evaluation: Compare output to set of reference captions

Flickr8k example image and captions

1. A man in street racer armor is examining the tire of another racers motor bike.
2. The two racers drove the white bike down the road.
3. Two motorists are riding along on their vehicle that is oddly designed and colored.
4. Two people are in a small race car driving by a green hill.
5. Two people in racing uniforms in a street car.

Automatic Image Captioning

Current Metrics: BLEU, Meteor, TERp

Relies on n-gram precision/recall, not necessarily robust to paraphrases, content selection variation, word order

MT metrics can underperform (Hodosh et al. 2014; Reiter and Belz 2009; Elliott and Keller 2014)

Solution: Word vectors! For better gradient similarity, especially for thematically related items (e.g., racer, car, horn, wheel)

Related Previous Work

Representation Based Translation Evaluation Metrics (Chen and Guo, 2014)

Approach: Build sentence-level vector and then calculate cosine similarity.

Three subsystems concatenated together:

1. One hot vector addition
2. Average word vectors
3. Socher-style autoencoder

Results

Full systems works best

Averaged word embeddings best single representation for WMT Out-of-English task

RAE vectors best single for WMT Into-English task

My approach

Averaged word vectors is problematic because some words will be noisy and contribute little semantic content

Weighting should help.

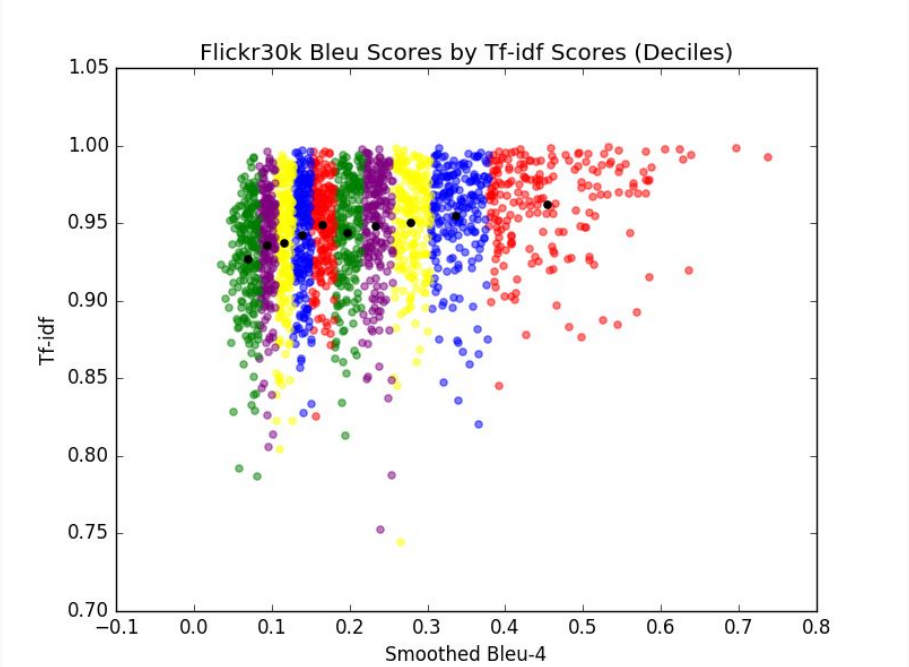
TF-IDF weight each word vector, then average.

Preliminary Results - ‘Gold’ data

2000 sample images from Flickr30k

Smoothed Bleu-4 deciles

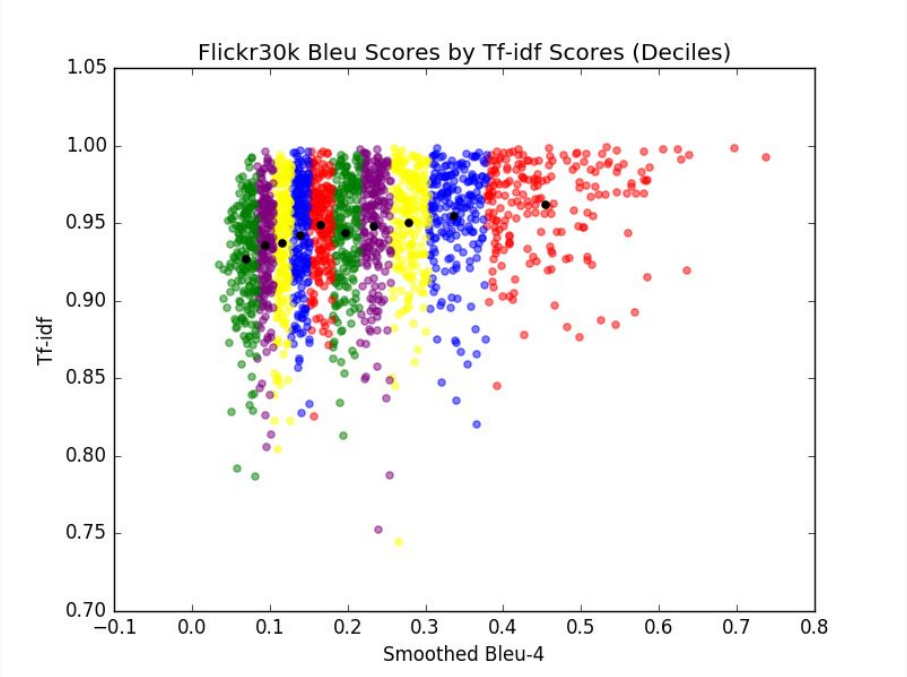
Black dots are averages



Preliminary Results - ‘Gold’ data

Good: Linear correlation between scores, tf-idf gets high scores

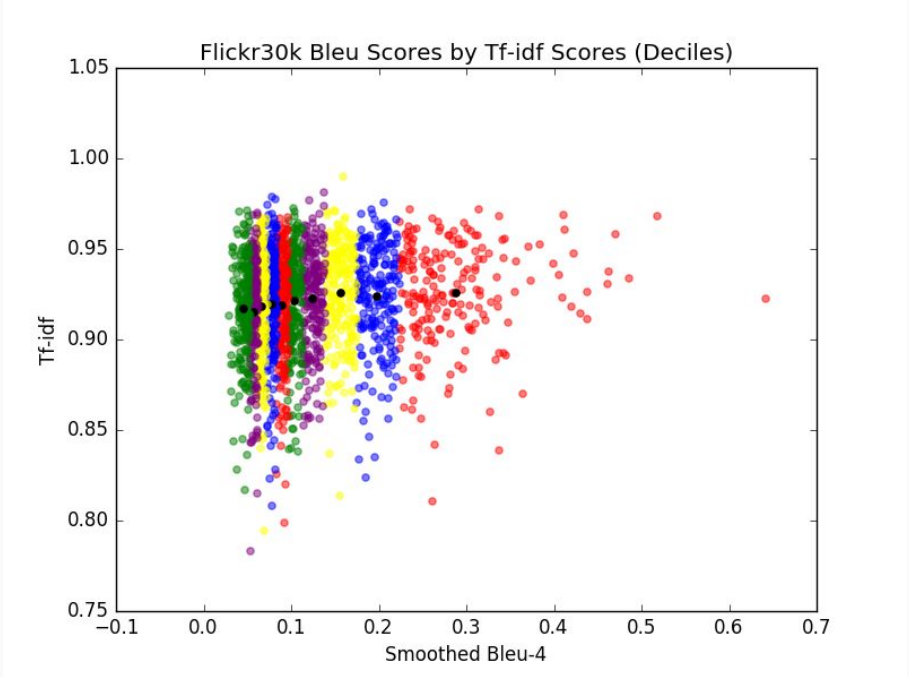
Bad: High variance?



Preliminary Results - 3 real, 2 random

Good: Lower top-end scores

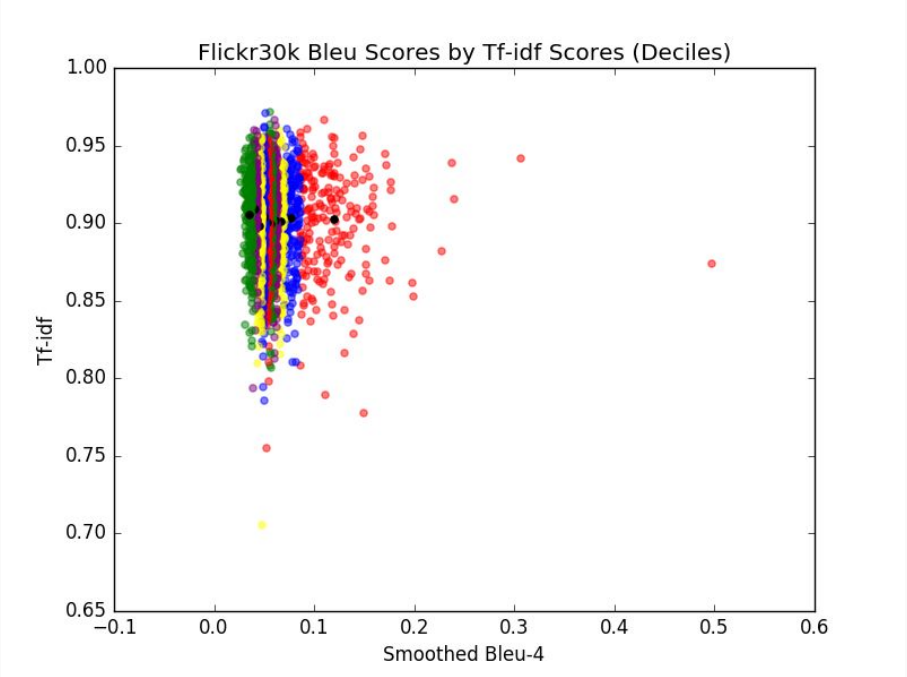
Bad: still pretty high?, averages in same low nineties range as ‘gold’ data



Preliminary Results - 5 random sentences

Good: Bleu is at floor

Bad: tf-idf still pretty high



Current Challenges

Evaluations:

- Ideally want human judgements, in order to correlate better than BLEU, Meteor, etc.
- Want to show robustness to paraphrase, restructuring, etc. but how to evaluate scores with different scales?
- Other evaluations?

Current Challenges

Existing metric:

Maybe tf-idf isn't ideal choice? Document here is defined as set of 5 captions. Simple frequency weighting better? (e.g., log probs from Gigaword)

Still insensitive to word order

Seems strangely high - maybe try unweighted, and other weighting schemes as sanity check

Current Challenges

Other suggestions?

Japanese Word Embeddings

花とは植物が成長してつけるものです。

Hanatowashokubutsugaseichoushitetsukerumonodesu.

flower.QUOT.TOP.plant.SUBJ.grow.do.attach.thing.COP

A flower is a thing that appears once a plant has grown.

Japanese Word Embeddings

Tokenization issue - no whitespace

花とは植物が成長してつけるものです。

Hanatowashokubutsugaseichoushitetsukerumonodesu.

Decisions here include status of case markers, postpositions, auxiliary verbs, etc.

What’s the best segmentation?

Probably depends on the task, but we can test different segmentations and find out!

Word-based (use dictionary)

Character-based (use string)

Other-based (use morpheme, POS, bunsetsu,?)

JUMAN

花とは植物が成長してつけるものです

花 か 花 名詞 6 普通名詞 1 * 0 * 0 "代表表記:花/か 漢字読み:音 カテゴリ:植物-部位"
と と と 助詞 9 格助詞 1 * 0 * 0 NIL
は は は 助詞 9 副助詞 2 * 0 * 0 NIL
植物 しょくぶつ 植物 名詞 6 普通名詞 1 * 0 * 0 "代表表記:植物/しょくぶつ カテゴリ:植物"
が が が 助詞 9 格助詞 1 * 0 * 0 NIL
成長 せいちょう 成長 名詞 6 サ変名詞 2 * 0 * 0 "代表表記:成長/せいちょう カテゴリ:抽象物"
して して する 動詞 2 * 0 サ変動詞 16 タ系連用テ形 14 "代表表記:する/する 付属動詞候補（基本） 自他動詞:自:成る/なる"
つける つける つける 動詞 2 * 0 母音動詞 1 基本形 2 "代表表記:付ける/つける 可能動詞:付く/つく 補文ト 付属動詞候補（基本） 自他動詞:自:付く/つく"
もの もの もの 名詞 6 形式名詞 8 * 0 * 0 NIL
です です だ 判定詞 4 * 0 判定詞 25 デス列基本形 27 NIL
EOS

http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

Previous Work

Morpheme-based embeddings (Socher et al. 2013)

It works good, try non-English languages for better results

Character-based embeddings (Utsumi 2014)

Mixed results. SVD. PPMI better than tf-idf weighting. Word-based methods still better unless rare words have frequent kanji. Also, novel words. Small window (n=1) good for synonymy, larger windows better for other relations

Best Tokenization for Vector Training?

Big question for agglutinative and polysynthetic languages. Could lead to better treatment of rare words, MWEs, etc.

Could be largely empirical, but what theoretical basis do we have for choosing one tokenization over another?

Potential Advantages of Character-based

Intuition: Many Japanese words consist of Kanji characters that have consistent meaning from compound to compound, like a morpheme:

違法 - ihou “illegal”

憲法 - kenpou “constitution”

家法 - kahou “family code/law”

Better generalization to unknown/undertrained words if they consist of Kanji we’ve seen before

target	context
ihou	hanzai
ihou	satsujin
hou	satsujin
hou	sousa
hou	kazoku

Evaluation

Utsumi 2014 evaluated word similarity by creating a 200 word synonym pair dataset. Score for task is minimum number of retrieved words that include a synonymous word for a target word. Can compare to his results.

I’ve generated word embeddings for the the JUMAN vs. JUMAN+kanji - just got Utsumi’s eval dataset recently and need to run it.

Suggestions for better evaluations/downstream tasks? Parsing, paraphrase, etc?