Towards Machine Translation of Homer

S. Sklaviadis   G. R. Crane

The Qualities of Literary Machine Translation , 2019

Outline

1. NLP in Ancient Greek
   - Introduction to the Corpus
2. Embeddings
   - Matrix factorization methods
   - canonical-GreekLit: Embedded
   - Evaluating embedding models
   - Digression: fitting GloVe on bitext
   - Dimension selection
   - Current Approaches
3. NMT
   - LSTMs with and without pre-trained embeddings
   - Loss plots
   - Decoded Iliad

Outline

- c. 100 million surviving words produced over more than 2,000 years (750 BCE-1453 CE)
- The canonical-GreekLit corpus: c. 10,000,000 words (`https://github.com/PerseusDL`)
- Homer's *Iliad* and *Odyssey* are two of the oldest Greek texts: c. 750 BCE
- Homer has been consistently translated from antiquity to the present

Outline

- The *Iliad* and *Odyssey* together are about 200,000 words.
- For this experiment we fitted 300 dimensional GloVe vectors.
  - Sampling is not a problem for this corpus size, and skipgram is otherwise equivalent.
- We compare the effect on NMT of 3 static, pretrained embedding models.
  - We vary the size and language of the embedding and use them to filter the input of a minimal LSTM NMT model.

Outline

canonical-GreekLit: Embedded

In [24]:
```
model_el.most_similar(positive=['Σωκράτης'], topn=20)
```

```
2019-08-18 10:42:46,714 : INFO : precomputing L2-norms of word weight vectors
```

Out[24]:
```
[('Εὐλογημένος', 0.5511736869812012),
 ('ἔφη·', 0.5490833520889282),
 ('ἔλεγεν', 0.5233275890350342),
 ('Σωφρονίσκου', 0.5107510089874268),
 ('Σιμωνίδης', 0.5060718059539795),
 ('Κάτων', 0.49643081426620483),
 ('Hamaker', 0.49119776487350464),
 ('Χαιρεφῶν.', 0.48857277631759644),
 ('Πλάτων', 0.4855552017688751),
 ('εἶπεν·', 0.4823710322380066),
 ('καρπαλίμως·', 0.48139774799346924),
 ('φιλοπτολέμοισιν', 0.46180281043052673),
 ('Δάμις', 0.460811972618103),
 ('Γουβάζης', 0.46022942662239075),
 ('Ξενοφῶν', 0.45998889207839966),
 ('Μυρτίλος', 0.4573240578174591),
 ('ὁ', 0.45698195695877075),
 ('Καῖσαρ,', 0.4552561342716217),
 ('Καλλίας', 0.4546433389186859),
 ('Ἀθηναῖος', 0.4501015543937683)]
```

Outline

```
# "boy" is to "father" as "girl" is to ...?
#model.most_similar(['girl', 'father'], ['boy'], topn=3)
#[('mother', 0.61849487), ('wife', 0.57972813), ('daughter', 0.56296098)]

model_el.most_similar(positive=['θεά', 'ἀνήρ'], negative=['θεός'], topn=20)

#γυναικῶν γυναῖκας (lemma: γυνή) αὐτῆς αὐτή θεά θεὰ
#ἄναξ ἀνὴρ ἀνήρ θεός
#NB women are referred to only in the plural in all of Iliad and Odyssey
```

```
2019-08-17 15:40:31,860 : INFO : precomputing L2-norms of word weight vectors
```

```
[('ἄειρεν', 0.5572900772094727),
 ('ἀλόχοισιν', 0.5487604141235352),
```

Analogy

Analogy

```
model_en_glove_vecmap_200.most_similar(positive=['goddess', 'man'], negative=['god'], topn=20)
```

```
('old', 0.622481107711792),
('spoke', 0.6113817095756531),
('Telemachus', 0.6093279719352722),
('Penelope', 0.6077815294265747),
('So', 0.6040306091308594),
('saying', 0.6017775535583496),
('woman', 0.5903725624084473),
```

A test from inflectional morphology

Tense stem variants should be more distant than mood variants:

Λύω

	PRESENT		FUTURE			AORIST			PERFECT	
	ACTIVE	MID.-PASS.	ACTIVE	MIDDLE	PASSIVE	ACTIVE	MIDDLE	PASSIVE	ACTIVE	MID.-PASS.
PRIMARY INDICATIVE	λύω λύεις λύει λύομεν λύετε λύουσι(ν)	λύομαι λύει/-ῃ λύεται λυόμεθα λύεσθε λύονται	λύσω λύσεις λύσει λύσομεν λύσετε λύσουσι(ν)	λύσομαι λύσει/-ῃ λύσεται λυσόμεθα λύσεσθε λύσονται	λυθήσομαι λυθήσει/-ῃ λυθήσεται λυθησόμεθα λυθήσεσθε λυθήσονται				λέλυκα λέλυκας λέλυκε λελύκαμεν λελύκατε λελύκασι(ν)	λέλυμαι λέλυσαι λέλυται λελύμεθα λέλυσθε λέλυνται
SECONDARY INDICATIVE	ἔλυον ἔλυες ἔλυε ἐλύομεν ἐλύετε ἔλυον	ἐλυόμην ἐλύου ἐλύετο ἐλυόμεθα ἐλύεσθε ἐλύοντο				ἔλυσα ἔλυσας ἔλυσε ἐλύσαμεν ἐλύσατε ἔλυσαν	ἐλυσάμην ἐλύσω ἐλύσατο ἐλυσάμεθα ἐλύσασθε ἐλύσαντο	ἐλύθην ἐλύθης ἐλύθη ἐλύθημεν ἐλύθητε ἐλύθησαν	ἐλελύκη ἐλελύκης ἐλελύκει ἐλελύκεμεν ἐλελύκετε ἐλελύκεσαν	ἐλελύμην ἐλέλυσο ἐλέλυτο ἐλελύμεθα ἐλέλυσθε ἐλέλυντο
SUBJUNCTIVE	λύω λύῃς λύῃ λύωμεν λύητε λύωσι(ν)	λύωμαι λύῃ λύηται λυώμεθα λύησθε λύωνται				λύσω λύσῃς λύσῃ λύσωμεν λύσητε λύσωσι(ν)	λύσωμαι λύσῃ λύσηται λυσώμεθα λύσησθε λύσωνται	λυθῶ λυθῇς λυθῇ λυθῶμεν λυθῆτε λυθῶσι(ν)	λελύκω λελύκῃς λελύκῃ λελύκωμεν λελύκητε λελύκωσι(ν)	λελυμένος ὦ etc.
OPTATIVE	λύοιμι λύοις λύοι λύοιμεν λύοιτε λύοιεν	λυοίμην λύοιο λύοιτο λυοίμεθα λύοισθε λύοιντο	λύσοιμι λύσοις λύσοι λύσοιμεν λύσοιτε λύσοιεν	λυσοίμην λύσοιο λύσοιτο λυσοίμεθα λύσοισθε λύσοιντο	λυθησοίμην λυθήσοιο λυθήσοιτο λυθησοίμεθα λυθήσοισθε λυθήσοιντο	λύσαιμι λύσαις (1) λύσαι (2) λύσαιμεν λύσαιτε λύσαιεν (3)	λυσαίμην λύσαιο λύσαιτο λυσαίμεθα λύσαισθε λύσαιντο	λυθείην λυθείης λυθείη λυθεῖμεν (4) λυθεῖτε (4) λυθεῖεν (4)	λελύκοιμι λελύκοις λελύκοι λελύκοιμεν λελύκοιτε λελύκοιεν	λελυμένος εἴην etc.
IMPERATIVE	λῦε λυέτω λύετε λυόντων	λύου λυέσθω λύεσθε λυέσθων				λῦσον λυσάτω λύσατε λυσάντων	λῦσαι λυσάσθω λύσασθε λυσάσθων	λύθητι λυθήτω λύθητε λυθέντων		λέλυσο λελύσθω λέλυσθε λελύσθων
PARTICIPLE M F N	λύων, -οντος λύουσα, -ης λῦον, -οντος	λυόμενος, λυομένη, λυόμενον	λύσων, -οντος λύσουσα, -ης λῦσον, -οντος	λυσόμενος, λυσομένη, λυσόμενον	λυθησόμενος λυθησομένη λυθησόμενον	λύσας, -αντος λύσασα, -σης λῦσαν, -αντος	λυσάμενος, λυσαμένη, λυσάμενον	λυθείς, -έντος λυθεῖσα, -σης λυθέν, -έντος	λελυκώς, -ότος λελυκυῖα, -ας λελυκός, -ότος	λελυμένος, λελυμένη, λελυμένον
INFINITIVE	λύειν	λύεσθαι	λύσειν	λύσεσθαι	λυθήσεσθαι	λῦσαι	λύσασθαι	λυθῆναι	λελυκέναι	λελύσθαι

(1) also: λύσειας; (2) also: λύσειε; (3) also: λύσειαν; (4) also: λυθείημεν, λυθείητε, λυθείησαν.
Nota Bene: λύσαι (optative), λῦσαι infinitive and imperative. Compare: παιδεύσαι (opt.), παίδευσαι (imp.), παιδεῦσαι (inf.)

1

[1] https://classics.uchicago.edu/files/luw.pdf

Matrix factorization embeddings

Not there yet...

```
1 model_el.relative_cosine_similarity('ποιεῖν', 'ποιεῖ')
```

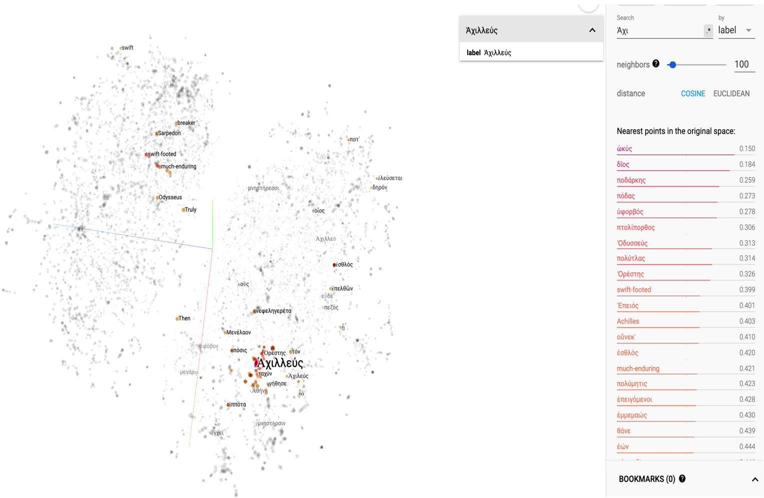
```
0.0713372947758188
```

- Infinitive vs. 3rd singular present primary indicative
- We can refine these models, however, neural embedding models are taking over quickly

Outline

Digression: Fitting GloVe on Bitext

A curious bag-of-words result



Outline

Dimension Selection

- Dimensionality of word embeddings has a big impact on their performance.
  - An embedding model with a small dimensionality is typically not expressive enough to capture all possible word relations, whereas one with a very large dimensionality suffers from over-fitting.
- Dimensionality is usually selected either ad hoc or by grid search.
- An empirical approach is to first train many embeddings of different dimensionalities, evaluate them on a functionality test (like word relatedness or word analogy), and pick the one with the best performance.

Dimension Selection

- For embedding algorithms that can be formulated as explicit or implicit matrix factorizations such as the LSA, skip-gram and GloVe, Yin & Shen (2018) propose a rigorous dimensionality selection procedure, by measuring the quality of the trained embeddings through a formal metric.
- Patel & Bhattacharyya (2018) propose lower bounds on the the number of dimensions, based on the number of pairwise equidistant words in the corpus vocabulary (as defined by some distance/similarity metric).
  - They suggest that going below these bounds results in degradation of quality of learned word embeddings.

Outline

Current Approaches

Recent developments in words embeddings include ELMo and BERT.

- In ELMo, each word is assigned a representation which is a function of the entire corpus' sentences.
  - The embeddings are computed from the internal states of a two-layers bidirectional Language Model.
- The relation between these newer Neural Network methods to methods that have been shown to be interpretable in terms of matrix factorization is still being explored.

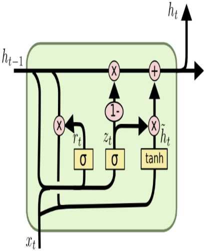
Outline

1. NLP in Ancient Greek
   - Introduction to the Corpus
2. Embeddings
   - Matrix factorization methods
   - canonical-GreekLit: Embedded
   - Evaluating embedding models
   - Digression: fitting GloVe on bitext
   - Dimension selection
   - Current Approaches
3. NMT
   - LSTMs with and without pre-trained embeddings
   - Loss plots
   - Decoded Iliad

Plain LSTM

- Plain LSTM: 7.2
- Homer in Greek: 6.0
- Homer bitext: 6.2
- canonnical-GreekLit: 6.4

2



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

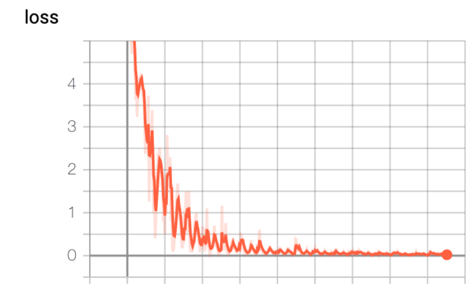
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

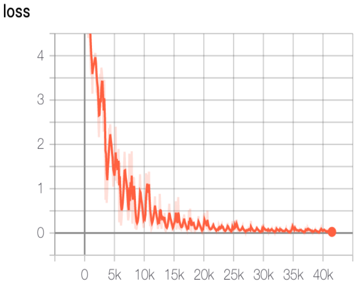
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

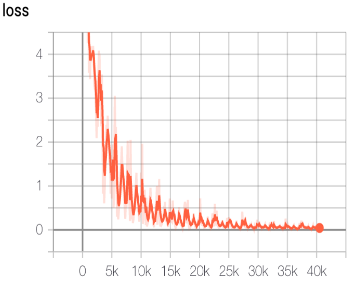
[2] OpenNMT, SacreBLEU, http://colah.github.io/posts/2015-08-Understanding-LSTMs/

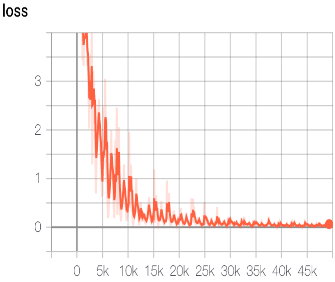
Outline

Loss Plots


loss
4
3
2
1
0


loss
4
3
2
1
0
0 5k 10k 15k 20k 25k 30k 35k 40k


loss
4
3
2
1
0
0 5k 10k 15k 20k 25k 30k 35k 40k


loss
3
2
1
0
0 5k 10k 15k 20k 25k 30k 35k 40k 45k

Outline

Decoded Iliad

It almost makes sense!

Plain LSTM

But if thou son of Peleus was first in the bidding of the first when thou hast question many horses and to catch battle if so be he mayest slay me from the Argives and Ilios let him fight even from Ilios and many handmaids layeth goodly goats and those son of Euaemon and upsprang Thoas best of the Achaeans and ever was it to fight against the gods with a bane What one came the daughter of Zeus and said

Butler, 1898

The wrath sing goddess of Peleus' son Achilles that destructive wrath which brought countless woes upon the Achaeans and sent forth to Hades many valiant souls of heroes and made them themselves spoil for dogs and every bird thus the plan of Zeus came to fulfillment from the time when first they parted in strife Atreus' son king of men and brilliant Achilles Who then of the gods was it that brought these two together to contend?