Evaluation: claims, evidence, significance

Sharon Goldwater

8 November 2019

School of informatics

First: Past exam question

HMMs are sometimes used for *chunking*: identifying short sequences of words (chunks) within a text that are relevant for a particular task. For example, if we want to identify all the person names in a text, we could train an HMM using annotated data similar to the following:

On/**O** Tuesday/**O** ,/**O** Mr/**B** Cameron/**I** met/**O** with/**O** Chancellor/**B** Angela/**I** Merkel/**I** ./**O**

There are three possible tags for each word: **B** marks the beginning (first word) of a chunk, **I** marks words inside a chunk, and **O** marks words outside a chunk. We also use **SS** and **ES** tags to mark the start and end of each sentence, respectively.

Crucially, the **O** and **SS** tags may not be followed by **I** because we need to have a **B** first indicating the beginning of a chunk.

What, if any, changes would you need to make to the Viterbi algorithm in order to use it for tagging sentences with this **BIO** scheme? How can you incorporate the constraint on which tags can follow which others?

Possible answers

Two students posted their own answers on Piazza (thank you!). For each one, does it get full credit, partial credit, or no credit? What is a 'full credit' answer?

Answer 1:

Change to the Viterbi: prevent the O and SS tags from following I. Incorporate the constraint on which tags can follow which others: make the transition probability equal zero to prevent some tags from following other tags.

Answer 2:

The Viterbi algorithm needs to be extended to dealing with trigrams and unknown words on this BIO scheme. Chancellor/B Angela/I Merkel/I is a trigram in the example sentence. And an unknown word started with a capitalized letter is more likely to be a proper noun.

Transition probabilities such as $P(*, O|I)$ and $P(*, SS|I)$ should be set to zero to incorporate transition constraints. Note that * is a wildcard that can match any from ES, SS, B, I, O

Discussion of Answer 1

Change to the Viterbi: prevent the O and SS tags from following I.

Incorporate the constraint on which tags can follow which others: make the transition probability equal zero to prevent some tags from following other tags.

Discussion of Answer 2

The Viterbi algorithm needs to be extended to dealing with trigrams and unknown words on this BIO scheme. Chancellor/B Angela/I Merkel/I is a trigram in the example sentence. And an unknown word started with a capitalized letter is more likely to be a proper noun.

Transition probabilities such as $P(*, O|I)$ and $P(*, SS|I)$ should be set to zero to incorporate transition constraints. Note that * is a wildcard that can match any from ES, SS, B, I, O

Today: Evaluation and scientific evidence

Throughout, we've discussed various evaluation measures and concepts:

- perplexity, precision, recall, accuracy
- comparing to baselines and toplines (oracles)
- using development and test sets
- extrinsic and intrinsic evaluation

Today: how do these relate to scientific hypotheses and claims? How should we state claims and evaluate other people's claims?

This, too, is an ethical issue

Scientific clarity and integrity are linked:

- Claims should be specific and appropriate to the evidence.
- Hypotheses cannot be "proved", only "supported".
- These days, claims are not just viewed by scientifically informed colleagues, but can make headlines...
- ...which means over-claiming can mislead the public as well as other researchers.

Just one recent example

Paper titled "Achieving Human Parity on Automatic Chinese to English News Translation" (Hassan et al., 2018).

- Headlines such as "Microsoft researchers match human levels in translating news from Chinese to English" (ZDNet, 14/03/18).
- On the bright side, at least they didn't call it "Achieving Human Parity with Machine Translation".
- But what is the real problem here?

As good as humans??

The problem: standard MT evaluation methods work on **isolated sentences**.

- Hassan et al. (2018) released their data/results (good!), allowing further analysis.
- Läubli et al. (2018) showed that when sentences were presented **in context**, human evaluators **did** prefer human translations.
- But of course this part does not make the news...

(Follow-up work proposes new test sets for targeted evaluation of discourse phenomena (Bawden et al., 2018))

Another (hypothetical) example

Student project compares existing parser (Baseline) to fancy new method (FNM).

- Uses Penn Treebank WSJ corpus, standard data splits.
- Develops and tunes FNM on development partition.
- F-scores after training on training partition:
  - Baseline: 91.3%, FNM: 91.8%.

Student concludes: "FNM is a better parser than Baseline."

1. Define the scope of the claim

- Experiment was run on a particular corpus of a particular language.
- We have no evidence beyond that.

More specific claim: "FNM is better at parsing Penn WSJ than Baseline."

- Conclusions/future work might say it is therefore worth testing on other corpora to see if it that claim generalizes beyond Penn WSJ.

2. Be specific: "better" how?

Depending on the situation, we might care about different aspects.

2. Be specific: "better" how?

Depending on the situation, we might care about different aspects.

- Accuracy? Speed of training? Speed at test time? Robustness? Interpretability?

More specific claim: "FNM parses Penn WSJ more accurately than Baseline."

Even better, include tradeoffs: "FNM parses Penn WSJ more accurately than Baseline, but takes about twice as long to parse each sentence."

3. Was the comparison fair?

Even a specific claim needs good evidence. In this case,

- Good: both systems trained and tested on the same data set.
- Possible problem: lots of tuning done for FNM, but apparently no tuning for Baseline. This is a common problem in many papers.
  - If Baseline was originally developed using some other corpus, FNM is tuned to this specific corpus while Baseline is not!
- Possible solutions:
  - spend equal effort tuning both systems
  - tune FNM on WSJ, then test both systems on some other corpus without re-tuning either. (This also provides a stronger test of generalization/robustness.)

4. Are the results statistically significant?

That is, are any differences real/meaningful or just due to random chance?

- Intuition: suppose I flipped a coin 20 times and got 13 heads. Is this sufficient evidence to conclude that my coin is not fair?
- Here, the randomness is due to the coin flip. Where is the randomness in NLP experiments?

1. Randomness due to data samples

- We evaluate systems on a **sample** of data. If we sampled differently, we might get slightly different results.
- "FNM is more accurate than Baseline on WSJ." How likely is this to be true in each case if we tested on N sentences from WSJ?

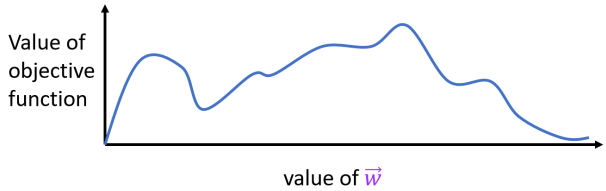
	Baseline F1	FNM F1
$N = 5$	73.2%	85.1%
$N = 500$	73.2%	85.1%
$N = 50000$	73.2%	85.1%
$N = 50000$	85.0%	85.1%

Does it matter?

- With a large enough sample size (test set), a 0.1% improvement might be a real difference. But still worth asking:
  - Is such a small difference enough to matter?
  - Especially since I usually care about performance on a range of data sets, not just this one.
- In many cases, our test sets are large enough that any difference that's enough to matter is also statistically significant.
- So, randomness of the data may not be our biggest worry...

2. Randomness in the algorithm

Some algorithms have non-convex objective functions: e.g., expectation-maximization, unsupervised logistic regression, multilayer neural networks.



So, different runs of the algorithm on the same data may return different results.

Now what to do?

- Where possible, run several times (different random seeds) and report the average and standard deviation.
- Some algorithms so time-consuming that this can be difficult (especially for many different settings).
- May be still possible to get a sense using a few settings.
- Either way, be cautious about claims!

A note about statistical significance

In some cases (small test sets, lots of variability between different runs) it may be worth running a **significance test**.

- Such tests return a **p-value**, where p is the probability of seeing a difference as big as you did, **if** there was actually no difference.
- Example: "FNM's F1 (87.5%) is higher than Baseline's (87.2%); the difference is significant with $p < 0.01$"
  - This means that the chance of seeing a difference of 0.3% would be less than 0.01 if FNM's and Baseline's performance is actually equivalent.

A note about statistical significance

In some cases (small test sets, lots of variability between different runs) it may be worth running a **significance test**.

- Lots of subtleties here, and also easy to over-trust (or mis-use or mis-understand) such tests.
- Please still consider whether your difference is enough to matter, even if real.

Evidence in Assignment 2

- Assignment 2 is intended mainly as an exploratory project, so you don't necessarily need to make definitive conclusions.
- But you should still be specific about any hypotheses/claims.
- We are looking for a start at thinking about some of the issues involved with distributional similarity, and a clearly written report.
- For some hints (mainly about correlation), please see the separate slides (also linked from today, but I will not go through them in lecture).

Summary

Good scientific practice involves making **claims** supported by **evidence**. To do this well,

- Claims should be specific, stating the aspects and scope for which you actually have evidence.
- Experiments should be designed for fair comparison.
- Results should be meaningful: not due to random chance, and large enough to matter.

References

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2018). Evaluating discourse phenomena in neural machine translation. In *16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., et al. (2018). Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Läubli, S., Sennrich, R., and Volk, M. (2018). Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796.