Project 2: Machine Translation with CRFs

May 3, 2017

In this project you will implement a probabilistic model for hierarchical machine translation somewhat based on the model of Blunsom et al. (2008).

In summary, your task will be to

- implement a latent variable model of translation formulated as a conditional random field;
- hypothesise a latent ITG tree (a synchronous binary tree) mapping between source and target;
- you will experiment with maximum likelihood estimation via stochastic gradient-based optimisation.

This project will help you familiarise yourself with

- bitext parsing
- hypergraphs
- inside-outside
- ancestral sampling
- undirected graphical models
- gradient-based optimisation

1 Bitext parsing

You will be provided with a constrained version of the following grammar, where Σ is the source vocabulary and Δ is the target vocabulary.[1]

$$
\begin{array}{lcll}
\mathrm{S} & \rightarrow & \mathrm{X} & \\
\mathrm{X} & \rightarrow & [\mathrm{X}\,\mathrm{X}] & \\
\mathrm{X} & \rightarrow & \langle \mathrm{X}\,\mathrm{X} \rangle & \\
\mathrm{X} & \rightarrow & x/y & \text{where } x \in \Sigma \text{ and } y \in \Delta \\
\mathrm{X} & \rightarrow & \epsilon/y & \text{where } y \in \Delta \\
\mathrm{X} & \rightarrow & x/\epsilon & \text{where } x \in \Sigma
\end{array}
$$

[1] We will constrain translation pairs to likely pairs under IBM model 1.

Your first task is to build a synchronous parser for such grammar (a variant of Earley parser), your parser should be general enough so that you can do the following

- Instantiate $\mathcal{D}(x) = \{d : \text{yield}_\Sigma(d) = x\}$, i.e., the set of derivations compatible with a source sentence x. Do you expect this to be a finite set? Why?
- Instantiate $\mathcal{D}(x, y) = \{d : \text{yield}_\Sigma(d) = x \text{ and } \text{yield}_\Delta(d) = y\}$, i.e., the set of derivations compatible with a source sentence x and its translation y. Do you expect this to be a finite set? Why?
- Instantiate $\mathcal{D}_n(x) = \{d : \text{yield}_\Sigma(d) = x \text{ and } |\text{yield}_\Delta(d)| \leq n\}$, i.e., the set of derivations that translates x to a target sentence no longer than n words. Do you expect this to be a finite set? Why?
- Compute the inside value of a node.
- Compute the outside value of a node.
- Find the best derivation and perform ancestral sampling.
- Optional (1 extra point): be able to instantiate $\mathcal{D}(x, \mathbf{y})$ where instead of a single reference y you parse a set of references $\mathbf{y}$

for these sets and operations will be crucial in fitting the CRF.

2 CRF

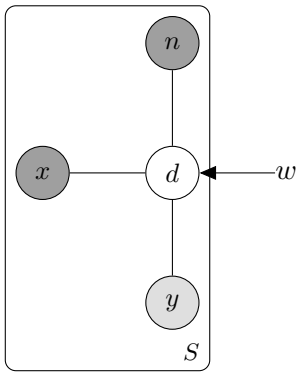

Figure 1: Conditional random field for translation with latent ITG trees.

2.1 Model

You will be working with a globally normalised model

$$
\begin{aligned}
P(y, d|x, n, w) &= \frac{\exp\left(w^\top \phi(x, y, d)\right)}{\sum_{y' \in \mathcal{Y}_n(x)} \sum_{d' \in \mathcal{D}_n(x, y')} \exp\left(w^\top \phi(x, y', d')\right)} \\
&= \frac{\exp\left(w^\top \phi(x, y, d)\right)}{\sum_{d' \in \mathcal{D}_n(x)} \exp\left(w^\top \phi(x, y', d')\right)} \quad \text{where } y' = \text{yield}_\Delta(d') \ .
\end{aligned} \tag{1}
$$

$\mathcal{D}_n(x)$ is determined by the ITG of Section 1, a derivation d fully determines a string pair (x, y), n is a length constraint, $w \in \mathbb{R}^d$ is a vector of parameters, and ϕ is a feature function mapping from $\mathcal{D}_n(x)$ to $\mathbb{R}^d$.[2]

We are going to use an edge-factored model, that is, our features will score independent steps (edges in a hypergraph) in ITG derivations. Because this is a conditional random field, we do not model the source sentence x, this means that global features of x are always available when scoring edges.

The minimum feature set is the following:

- word pair (for translation/deletion/insertion rules);
- rule indicator (e.g. translation, deletion, insertion, monotone, inverted);
- source inside features (e.g. average embedding or bidirectional representation);
- source outside features (e.g. average embedding or bidirectional representation);
- source span features (e.g. length);
- target span features (e.g. length);
- source skip-bigrams.

2.2 Training

The likelihood of an observation (x, y) is given by marginalisation of latent derivations

$$
P(y|x, n, w) = \sum_{d \in \mathcal{D}_n(x)} P(y, d|x, n, w) \ . \tag{2}
$$

[2]The length constraint n is necessary in order to guarantee that the normaliser in Equation (1) is finite for an arbitrary choice of parameters w.

For a given length constraint n, the parameters of the model can be optimised for maximum likelihood via gradient-based optimisation

$$\nabla_w \mathcal{L}(w|x, y, n) = \mathbb{E}_{P_{D|x,y,n,w}}[\phi(X, Y, D)] - \mathbb{E}_{P_{Y,D|x,n,w}}[\phi(X, Y, D)] \ , \tag{3}$$

where $\mathcal{L}(w|x, y, n) = \log P(y|x, n, w)$ is the log-likelihood of the observations.

In this part your first task is to prove Equation (3). Then, you will implement the CRF specified in (1) and optimise its parameters with stochastic optimisation

$$w^{(t+1)} = w^{(t)} + \gamma^{(t)} \nabla_{w^{(t)}} \mathcal{L}(w^{(t)}|x, y, n) \ . \tag{4}$$

You should experiment with and without an L_2 regulariser.[3] We will make available a validation set which you can use to tune the regulariser strength. Also note that evaluating the likelihood (2) for the complete training data at every epoch may be too expensive, thus you may perform model selection purely based on likelihood of validation data.[4]

2.3 Prediction

You will experiment with 3 decision rules.

Viterbi decoding Finding the best translation

$$\begin{aligned} y^\star &= \arg\max_{y \in \mathcal{Y}_n(x)} P(y|x) \\ &= \arg\max_{y \in \mathcal{Y}_n(x)} \sum_{d \in \mathcal{D}_n(x,y)} P(y, d|x) \end{aligned} \tag{5}$$

is intractable due to the marginalisation of latent derivations. A common approximation is to search for the highest scoring derivation instead.

$$y^\star = \text{yield}_\Delta \left\{ \arg\max_{d \in \mathcal{D}_n(x)} P(y, d|x) \right\} \tag{6}$$

Marginal decoding The marginal probabilities in (5) can be approximated by ancestral sampling, where the arg max is taken with respect to a sample rather than the complete space.

[3] An L_2 regulariser is an approximation to a Gaussian prior on the parameter vector.

[4] In principle, one could perform model selection using validation BLEU, but that would require a decision rule for prediction which will investigate in Section 2.3.

Minimum Bayes risk decoding We may introduce a loss $l(h, r)$ that compares a hypothesis h and a reference r (e.g. $1-\text{BLEU}(h, r)$) and make a decision based on minimum risk

$$\begin{aligned} y^\star &= \arg\min_{y' \in \mathcal{Y}_n(x)} \mathbb{E}_{P_{Y|x,n,w}}[l(y', y)] \\ &\approx \arg\min_{y' \in \bar{\mathcal{Y}}_n(x)} \text{yield}_\Delta \left\{ \sum_{i=1}^{N} l(y', \text{yield}_\Delta(d_i)) \right\} \end{aligned} \tag{7}$$

where $d_i \sim P(Y, D|x, n, w)$ is a sample and $\bar{\mathcal{Y}}_n(x)$ is the set of sampled candidates.

NOTE ON ANNEALING Optimisation through sampling often requires annealing, where a positive real number α controls how flat or peaked a distribution is.

$$P_\alpha(y, d|x, n, w) = P(y, d|x, n, w)^\alpha \propto \exp\Big(\alpha \times w^\top \phi(x, y, d)\Big) \tag{8}$$

After training you can tune α for best validation BLEU.

3 Data and code

Chinese-English (BTEC) data for training/validation/test (Takezawa et al., 2002).

- sentences up to 30 words
- multiple references for validation
- translation lexicon from IBM model 1

We have prepared python classes that implement a basic ITG parser, check our course page.

4 Report

You should use latex for your report, and you should use the ACL template available from `http://acl2017.org/downloads/acl17-latex.zip` (unlike the template suggests, your submission should not be anonymous).

We expect short reports (4 pages plus references) written in English. The typical submission is organised as follows:

- abstract: conveys scope and contributions;
- introduction: present the problem and relevant background;
- model: technical description of models;
- experiments: details about the data, experimental setup and findings;
- conclusion: a critical take on contributions and limitations.

5 Submission

You should submit a tgz file containing a folder (folder name `lastname1.lastname2`) with the following content:

- Test predictions (one translation per line) using your best model
- Report as a single pdf file (filename: `report.pdf`)

Your report may contain a link to an open-source repository (such as github), but please do not attach code or additional data to your tgz submission.

You can complete your project submission on Blackboard no later than **May 19, 23:59 GMT-8**.

6 Assessment

Your report will be assessed by two independent reviewers according to the following evaluation criteria:

1. Scope (max 2 points): Is the problem well presented? Do students understand the challenges/contributions?
2. Theoretical description (max 3 points): Are the models presented clearly and correctly?
3. Empirical evaluation (max 3 points): Is the experimental setup sound/convincing? Are experimental findings presented in an organised and effective manner?
4. Writing style (max 2 points): use of latex, structure of report, use of tables/figures/plots, command of English.
5. Extra (max 1 point).

References

Blunsom, P., Cohn, T., and Osborne, M. (2008). A discriminative latent variable model for statistical machine translation. In *Proceedings of ACL-08: HLT*, pages 200–208, Columbus, Ohio. Association for Computational Linguistics.

Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H., and Yamamoto, S. (2002). Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Third International Conference on Language Resources and Evaluation*, LREC, Las Palmas, Canary Islands - Spain. European Language Resources Association.