Concept Map

CSE 713

Group Task 3

20166006, Tanzim Ahmed
20166019, Nurun Nahar
20266006, Maksudur Rahman Sohag
18101496, Mirza Ahmad Shayer
21141046, Afnan Ahmed Crystal

Speech and Language Processing (3rd ed. draft)
Dan Jurafsky and James H. Martin

Chapter 11
Machine Translation

Chapter 11 : Machine Translation and Encoder-Decoder Models

This chapter introduces machine translation (MT), the use of computers to translate from one language to another.
Of course translation, in its full generality, such as the translation of literature, or poetry, is a difficult, fascinating, and intensely human endeavor, as rich as any otherarea of human creativity.

11.1 Language Divergences and Typology

Some aspects of human language seem to be universal, holding true for every language, or are statistical universals, holding true for most languages. Many universals arise from the functional role of language as a communicative system by humans

11.1.1 Word Order Typology

As we hinted it in our example above comparing English and Japanese, languages differ in the basic word order of verbs, subjects, and objects in simple declarative clauses.

11.1.2 Lexical Divergences

For any translation, the appropriate word can vary depending on the context. The English source-language word bass. Sometimes one language places more grammatical constraints on word choice than another.
The way that languages differ in lexically dividing up conceptual space may be more complex than this one-to-many translation problem, leading to many-to-many mappings.

11.1.3 Morphological Typology

Morphologically, languages are often characterized along two dimensions of variation. The first is the number of morphemes per word, ranging from isolating languages like Vietnamese and Cantonese, in which each word generally has one morpheme, to polysynthetic languages like Siberian Yupik (“Eskimo”), in which a single word may have very many morphemes,

11.1.4 Referential density

Finally, languages vary along a typological dimension related to the things they tend to omit. Some languages, like English, require that we use an explicit pronoun when talking about a referent that is given in the discourse.
Languages that can omit pronouns are called pro-drop languages. Even among the pro-drop languages, there are marked differences in frequencies of omission.

11.2 The Encoder-Decoder Model

Encoder-decoder networks, or sequence-to-sequence networks, are models capable of generating contextually appropriate, arbitrary length, output sequences.
Encoder-decoder networks have been applied to a very wide range of applications including machine translation, summarization, question answering, and dialogue.

11.3 Encoder-Decoder with RNNs

Translating a single sentence (inference time) in the basic RNN version of encoder-decoder approach to machine translation. Source and target sentences are concatenated with a separator token in between, and the decoder uses context information from the encoder’s last hidden state.

11.3.1 Training the Encoder-Decoder Model

Encoder-decoder architectures are trained end-to-end, just as with the RNN language models.
Each training example is a tuple of paired strings, a source, and a target. Concatenated with a separator token, these source-target pairs can now serve as training data.
For MT, the training data typically consists of sets of sentences and their translations.

11.4 Attention

The simplicity of the encoder-decoder model is its clean separation of the encoder which builds a representation of the source text from the decoder, which uses this context to generate a target text. This final hidden state is thus acting as a bottleneck: it must represent absolutely everything about the meaning of the source text, since the only thing the decoder knows about the source text is what’s in this context vector.
The attention mechanism is a solution to the bottleneck problem, a way of allowing the decoder to get information from all the hidden states of the encoder, not just the last hidden state.

11.5 Beam Search

Indeed, greedy search is not optimal, and may not find the highest probability translation.
Instead, decoding in MT and other sequence generation problems generally uses a method called beam search. In beam search, instead of choosing the best token to generate at each timestep, we keep k possible tokens at each step. This fixed-size memory footprint k is called the beam width, on the metaphor of a flashlight beam that can be parameterized to be wider or narrower.

11.6 Encoder-Decoder with Transformers

Scoring for beam search decoding with a beam width of k = 2. We maintain the log probability of each hypothesis in the beam by incrementally adding the logprob of generating each next token. Only the top k paths are extended to the next step

11.7 Some practical details on building MT systems

Machine translation models are trained on a parallel corpus, sometimes called a bitext, a text that appears in two (or more) languages.

11.7.1 Tokenization

Generally a shared vocabulary is used for the source and target languages, which makes it easy to copy tokens (like names) from source to target, so we build the wordpiece/BPE lexicon on a corpus that contains both source and target language data. Wordpieces use a special symbol at the beginning of each token; here’s a resulting tokenization from the Google MT system

11.7.2 MT corpora

Machine translation models are trained on a parallel corpus, sometimes called a bitext, a text that appears in two (or more) languages. Large numbers of parallel corpora are available. Standard training corpora for MT come as aligned pairs of sentences. When creating new corpora, for example for underresourced languages or new domains, these sentence alignments must be created.

11.7.3 Backtranslation

Backtranslation is a way of making use of monolingual corpora in the target language by creating synthetic bitexts. In backtranslation, we train an intermediate target-to-source MT system on the small bitext to translate the monolingual target data to the source language

11.8 MT Evaluation

Translations can be evaluated along two dimensions, adequacy and fluency.
adequacy: how well the translation captures the exact meaning of the source sentence. Sometimes called faithfulness or fidelity.
fluency: how fluent the translation is in the target language (is it grammatical, clear, readable, natural).

11.8.1 Using Human Raters to Evaluate MT

An alternative is to do ranking: give the raters a pair of candidate translations, and ask them which one they prefer.
While humans produce the best evaluations of machine translation output, running a human evaluation can be time consuming and expensive.

11.8.2 Automatic Evaluation: BLEU

Consider a test set from a parallel corpus, in which each source sentence has both a gold human target translation and a candidate MT translation we’d like to evaluate. The BLEU metric ranks each MT target sentence by function of the number of n-gram overlaps with the human translation.
BLEU is actually not a score for a single sentence; it’s a score for an entire corpus of candidate translation sentences. More formally, the BLEU score for a corpus of candidate translation sentences is a function of the n-gram precision over all the sentences combined with a brevity penalty computed over the corpus as a whole

11.8.3 Automatic Evaluation: Embedding-Based Methods

The BLEU metric is based on measuring the exact word or n-grams a human reference and candidate machine translation have in common. However, this criterion is overly strict, since a good translation may use alternate words or paraphrases. A solution pioneered in early metrics like METEOR was to allow synonyms to match between the reference x and candidate x̃. More recent metrics use BERT or other embeddings to implement this intuition.

11.9 Bias Given and Ethical Issues

MT systems can be used in urgent situations where human translators may be unavailable or delayed: in medical domains, to help translate when patients and doctors don’t speak the same language, or in legal domains, to help judges or lawyers communicate with witnesses or defendants. In order to ‘do no harm’, systems need ways to assign confidence values to candidate translations, so they can abstain from giving incorrect translations that may cause harm.

11.10 Summary

Machine translation is one of the most widely used applications of NLP, and the encoder-decoder model, first developed for MT is a key tool that has applications throughout NLP.
Languages have divergences, both structural and lexical, that make translation difficult.
Human evaluation is the gold standard, but automatic evaluation metrics like BLEU, which measure word or n-gram overlap with human translations, or more recent metrics based on embedding similarity, are also commonly used.