16 Applications 1: Monolingual Sequence-to-sequence Problems

Up until now, we have largely used machine translation as an example of sequence-to-sequence learning tasks. However, as mentioned at the beginning of the course, sequence-to-sequence models are quite general, and can be used for a large number of tasks. There are also a number of other sequence-to-sequence tasks, and describes some of the unique features that make these tasks difficult or different from machine translation.

In this chapter we'll give some examples of sequence-to-sequence transduction tasks that are performed within a single language, translating, for example, English into English.

16.1 Paraphrase Generation

The most general form of translation between two sentences in the same language is paraphrasing: re-wording sentences into other sentences with the same content but different surface features. This technology has a number of applications including query expansion for information retrieval [30] or improving robustness of machine translation to lexical variations [3], and a few other specific applications described later.

Formally, in paraphrasing, we receive an input F and want to output a sentence E in the same language that has the same content but different wording. There are a few interesting features of paraphrasing (that also carry over to most monolingual transduction tasks) that make it more difficult (in some ways) and less difficult (in other ways) than machine translation between languages. The first difficulty is in the **task definition**; the question of "what is a paraphrase?" is not well defined and must be chosen appropriately to fit whatever downstream use case of paraphrasing is envisioned. One way to define paraphrasing is *bi-directional entailment*, where given two sentences F and E, F must be true if E is and vice-versa. However, it is quite unlikely that two sentences with different wording will have *exactly* the same meaning, as they will often differ in small nuances. Thus it may become necessary to relax this definition to allow any interesting or useful paraphrases to use as training or test data. For example, in the Microsoft Research Paraphrasing Corpus (MRPC; [9]), one of the early datasets of sentential paraphrases, use the rather loose definition of "mostly bidirectional entailment," which allows it to pick up the following pair of sentences:

> Charles O. Prince, 53, was named as Mr. Weill's successor.
> Mr. Weill's longtime confidant, Charles O. Prince, 53, was named as his successor.

However, this definition will not necessarily satisfy the needs of all paraphrasing tasks (or monolingual translation tasks in general), and the tasks described below have their own definitions which we'll cover in turn.

A second difficulty is **paucity of data**; unlike machine translation where relatively large corpora of bilingual text containing inputs F and outputs E are easy to come by, it is quite difficult to find large corpora of parallel text in the same language with the same meaning. There are a few examples of large-scale datasets with paraphrases, such as the Quora question pair dataset[52] and the MSCOCO captions dataset,[53] but these are rare and only exist for a

[52] `https://www.kaggle.com/c/quora-question-pairs`

[53] `http://cocodataset.org`

small number of limited domains and languages. Thus, it is generally necessary to train paraphrasing systems without large parallel resources, and thus some other source of information about which words and structures can be translated into each-other needs to be used. In general, there are two major methods for doing so: those based on *distributional similarity* and those based on *bilingual pivoting*.

Distributional similarity methods are based on the concept that words that appear in similar contexts tend to to be similar, much like the methods that are used to train word embeddings mentioned in Section 5. A first attempt at finding paraphrasable words and phrases based on distributional similarity is **discovering inference rules in text** (DIRT; [18]), which first uses a dependency parser to analyze sentences, then extracts paths through the dependency tree with empty "slots" that can be filled in by other words. These may take the shape of "X finds a solution to Y" or "X solves Y". Then out of the large number of paths extracted from a monolingual corpus, the method calculates the similarities in the distributions between the words that fill slot X and slot Y, and patterns where the distributions of X and Y are both similar are deemed as likely paraphrases.

One major problem with distributional similarity based paraphrase methods is that they do not have enough information to distinguish between distributionally similar but semantically different words. A stereotypical example of this is antonyms such as "love" and "hate", which often tend to occur in the same context. Another difficulty with distributional similarity based methods is that they are extremely sensitive to data sparsity: if a particular word or pattern only occurs one or a couple of times in a corpus then there is not enough information to disambiguate from other inputs. One method that has been highly effective in overcoming this problem and improving the quality of paraphrasing as a whole is the use of *bilingual data to learn monolingual paraphrases*. The idea behind these methods is simple: because words that get translated the same way in another language tend to have the same meaning, we can use information about how words are translated to find synonyms or synonymous phrases.

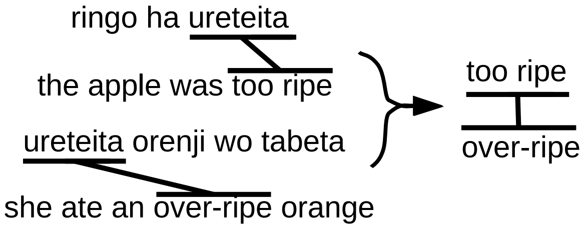

Figure 54: An example of extracting monolingual paraphrases from bilingual phrases.

For example, [1] describe a simple method to extract phrasal paraphrase candidates from bilingual machine translation training data using methods from phrase-based machine translation (Section 14) and pivoting, as shown in Figure 54. Basically, the idea is that we can calculate the probability of a paraphrase between English phrases $P(\boldsymbol{e}_2 \mid \boldsymbol{e}_1)$ by marginalizing over the probability of phrases in the source language:

$$P(\boldsymbol{e}_2 \mid \boldsymbol{e}_1) = \sum_{\boldsymbol{f}} P(\boldsymbol{e}_2 \mid \boldsymbol{f}) P(\boldsymbol{f} \mid \boldsymbol{e}_1). \tag{165}$$

This means that if we can extract a phrase table from a parallel text, as described in Section 14, we can build a paraphrasing model with no annotated monolingual text. This overall paradigm has proven quite effective, and is now the basis for the widely used paraphrase database PPDB [12][54], which contains paraphrases of words, phrases, and syntactic structures.

Once these paraphrasing rules have been extracted, they can be used in a number of ways. For example, they can be used in the phrase table or rule table of phrase-based machine translation systems to be described in Section 14, making it possible to calculate a translation probability $P(E \mid F)$ or $P(F \mid E)$, which can be combined with a language model $P(E)$ to generate both faithful and fluent paraphrases. These methods have also been used to train neural paraphrase *identification* models, where the neural model has to decide between rules that exist in a paraphrase table and those that do not [32]. However, there are few examples of neural paraphrase *generation* models that have been trained in such an unsupervised way, and they mostly are applied in the context of style transformation, which will be described in the next section.

One final difficult aspect of paraphrase generation is **how to evaluate** the generated paraphrases. One way to do so is to prepare some reference "correct" paraphrases, and measure BLEU score with respect to them, but when simply using this metric trivial solution of copying the source sentence as-is and treating it as a "paraphrase" will be an extremely difficult baseline to beat. However, while we would like the paraphrase to be accurate and fluent, we also need to ensure that they need to be significantly different from the original text. One example of an evaluation measure that considers this is PINC [6], which is like BLEU but considers not only the BLEU score, but also the dissimilarity from the original input.

Interested readers can find an extensive survey of paraphrasing in [19] or on `http://paraphrasing.org` (the latter is more up-to-date).

16.2 Style Transformation

A second variety of monolingual text transduction is *style transformation* or *style transfer*, which attempt to take a source sentence F and convert it into a sentence E in the same language with the same semantic content, but with a different style or register. These methods have been used in a number of different contexts:

Text Simplification: Conversion of text from a more complicated form to a less complicated one [5, 28]. This variety of transformation, which largely consists of simplifying syntax and replacing more difficult words for simpler ones, is particularly useful for second-language reading comprehension.

Register Conversion: "Register" is the type of language used in a particular setting, and conversion of register converts between these types of language. For example, it is possible to take the more informal text and convert it into more formal text appropriate for writing in business situations or meeting transcripts [22, 25]. Another example is converting offensive language into non-offensive language [31, 23].

Personal Style Conversion: It is also possible to convert between personal styles, taking text written in a neutral style and imbuing it with the traits of a particular author, such as literary figures such as Shakespeare [33] or cartoon characters [20].

[54] `http://paraphrase.org/`

Demographics-level Conversion: It is also possible to convert between the typical speaking style of particular demographic groups, such as male and female speakers or place of birth, etc. [26, 24].

Style transformation, in a way, is a strictly more difficult problem than paraphrasing by definition: in paraphrasing we need to generate an arbitrary output that has the same semantic content while in style transformation we additionally need the output to satisfy particular features of being simpler, more polite, or representative of a speaker or demographic group.

The simplest method for style transformation, if possible, is to create a large parallel corpus and use it to train a supervised model. This can be done in a limited number of situations. For example, for parliamentary proceedings such as the European Parliament or the Japanese Diet, it is possible to get both faithful transcripts of speech as it is actually spoken, and then also the text that actually appears in the parliamentary proceedings [22]. For famous books, such as the works of Shakespeare or the Bible, it is also common to be able to get versions in different register, or translations by different authors, which can provide a rich parallel training set [33]. However, for the great majority of tasks it is difficult to get parallel aligned corpora in the source and target styles, and thus unsupervised methods are required.

Because phrase-based translation models consist of both a translation model (which considers source F and target E) and a language model (which considers target E only), it is relatively easy to tailor them to the task of style transformation. Specifically, because we can assume that we have a large amount of data in the target style, we can train the language model E on only the target text, and use a large general-purpose paraphrase database such as PPDB as the translation model. One complication of this method is when large general-purpose paraphrase databases (such as PPDB) cannot cover specific phrases that are specific to the style to which we want to translate. In this case, [20] find that combining bilingual and distributional-similarity-based methods for obtaining paraphrases can improve coverage.

Within neural models, it is slightly harder to apply these to the task of unsupervised style transfer, as they generally model the conditional probability $P(E \mid F)$ directly, without incorporating a language model. One popular method for attempting to get around this restriction is through the use of cycle-based training [29], which is similar to methods for semi-supervised training for standard machine translation [7]. The basic idea behind these methods is that you first generate a hypothesis in the forward direction:

$$\hat{E} \sim P(\tilde{E} \mid F) \tag{166}$$

then calculate the probability of generating *the original sentence* given this sample:

$$P(F \mid \tilde{E}), \tag{167}$$

the intuition being that a $\tilde{E}$ where this conditional probability is high probably maintains the content of F better than a $\tilde{E}$ where the probability is low. This can be plugged into methods for optimizing sequence-to-sequence models based on arbitrary rewards such as reinforcement learning or minimum risk training, which will be covered more extensively in Section 18. Importantly for style transformation, because we would also like to have the output text be in an arbitrary style, we can add an additional loss function $\ell_{textstyle}(E)$ which gives a penalty to sentences that seem they are not in the appropriate style. This type of training is akin to *generative adversarial networks* [13], where this loss function is calculated based on a discriminative model that tries to guess the style of the generated output, and the model is given a penalty any time the output is judged to not be in the correct style. [34] propose a method where this discriminator can be trained using a language model $P(E)$, which makes it possible to efficiently train the discriminator over large-scale data before training the translation model itself.

[24] propose an alternative method for performing style transformation that, similarly to the bilingual paraphrase extraction methods above, takes advantage of the fact that we can get large amounts of bilingual text. This method works by pivoting through another language: first translating the input sentence F into a sentence in another language G, then translating back into the original language sentence E. In this case, $P(G \mid F)$ can be trained on all bilingual data between the source and pivot languages, and $P(E \mid G)$ can be trained on a subset of the parallel data that contains a particular stylistic trait.

16.3 Summarization

A final typical example of monolingual transduction tasks is **text summarization**. In the summarization task, compared to the methods above, the content differs between the source and the target: a larger body of text F is converted into a smaller amount of text E containing the *most salient information* in F for browsing purposes. This can be done at a number of levels:

Sentence Compression: The problem of compressing a single sentence into a shorter single sentence [16].

Single-document Summarization: The problem of compressing a single document into a shorter summary [4].

Multi-document Summarization: the problem of reducing the information in multiple documents into a single summary [2].

There are also typically two types of summarization: **extractive summarization** and **abstractive summarization**. In extractive summarization, we simply choose some content (usually one sentence at a time), and add these to the summary. In contrast, in abstractive summarization we actually generate a new summary, and systems using this approach have been created using the sequence-to-sequence models introduced in this course.

One unique element of summarization is that it is largely concerned with removing irrelevant content. Thus, many attempts, both using non-neural statistical systems and neural systems, focus on simply deleting words [16, 10]. In particular, tree-based methods that explicitly use syntax have found some favor, as this is a natural way to model that fact that we can "chop off" irrelevant phrases without a major change in the main content [21]. It is common to frame these problems as a constrained optimization problem; we want to delete words to achieve a summary with a certain length while maximizing the amount of relevant content that remains in the summary.

There have also been a number of methods that move beyond only deletion, and frame the problem as a sequence-to-sequence transduction problem. Successful methods have used tree substitution grammars [8], and attentional neural networks [27]. These models can be equipped with special mechanisms to copy words [14], or control the length of the summary [15].

Summarization systems are generally evaluated based on the amount of recall of important information that can be achieved within the limited summary length. The standard measure is ROUGE, which measures recall over n-grams [17], and it is also common to perform manual human evaluation as well.

Interested readers can find a more complete survey in [11].

16.4 Exercise

A potential exercise for this section would be to find and download a data set for one of these tasks, and run your sequence-to-sequence model on it and observe the results.

References

[1] Colin Bannard and Chris Callison-Burch. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 597–604, 2005.

[2] Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 550–557, 1999.

[3] Chris Callison-Burch, Philipp Koehn, and Miles Osborne. Improved statistical machine translation using paraphrases. In *Proceedings of the 2006 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 17–24, 2006.

[4] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. pages 335–336. ACM, 1998.

[5] Raman Chandrasekar, Christine Doran, and Bangalore Srinivas. Motivations and methods for text simplification. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, pages 1041–1044. Association for Computational Linguistics, 1996.

[6] David Chen and William Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 190–200, 2011.

[7] Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1965–1974, Berlin, Germany, August 2016. Association for Computational Linguistics.

[8] Trevor Cohn and Mirella Lapata. Sentence compression beyond word deletion. In *Proceedings of the 22th International Conference on Computational Linguistics (COLING)*, pages 137–144, 2008.

[9] William B Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005.

[10] Katja Filippova, Enrique Alfonseca, Carlos A. Colmenares, Lukasz Kaiser, and Oriol Vinyals. Sentence compression by deletion with lstms. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 360–368, 2015.

[11] Mahak Gambhir and Vishal Gupta. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66, 2017.

[12] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 758–764, 2013.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014.

[14] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1631–1640, 2016.

[15] Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[16] Kevin Knight and Daniel Marcu. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107, 2002.

[17] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, 2004.

[18] Dekang Lin and Patrick Pantel. Dirt – discovery of inference rules from text. pages 323–328. ACM, 2001.

[19] Nitin Madnani and Bonnie J. Dorr. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, pages 341–387, 2010.

[20] Masahiro Mizukami, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. Linguistic individuality transformation for spoken language. In *Proceedings of the International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, 2015.

[21] Hajime Morita, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. Subtree extractive summarization via submodular maximization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1023–1032, 2013.

[22] Graham Neubig, Shinsuke Mori, and Tatsuya Kawahara. A WFST-based log-linear framework for speaking-style transformation. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (InterSpeech)*, pages 1495–1498, 2009.

[23] Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 189–194, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[24] Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 866–876, 2018.

[25] Sudha Rao and Joel Tetreault. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, 2018.

[26] Sravana Reddy and Kevin Knight. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26, Austin, Texas, November 2016. Association for Computational Linguistics.

[27] Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 379–389, 2015.

[28] Matthew Shardlow. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1), 2014.

[29] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS)*, pages 6830–6841, 2017.

[30] Karen Sparck Jones and John I Tait. Automatic search term variant generation. *Journal of Documentation*, 40(1):50–66, 1984.

[31] Rajan Vaish and Andrés Monroy-Hernández. Crowdtone: Crowd-powered tone feedback and improvement system for emails. *arXiv preprint arXiv:1701.01793*, 2017.

[32] John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. Towards universal paraphrastic sentence embeddings. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[33] Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. Paraphrasing for style. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 2899–2914, 2012.

[34] Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. Unsupervised text style transfer using language models as discriminators. In *Proceedings of the 32nd Annual Conference on Neural Information Processing Systems (NIPS)*, 2018.