SMT Solutions

Question 1

1(a) Any statistical approach to MT requires the availability of aligned bilingual corpora which are (i) large, (ii) good-quality, and (iii) representative. Explain why all three requirements are important. What are some of the potential problems if any one of these requirements are not met?

If the corpora are small, or of poor quality, or are unrepresentative, then our statistical language models will be poor, so any results we achieve will be poor.
+
(i) Large:
We need a large aligned bilingual corpus so as to be able to have a better chance of achieving high accuracy translations by creating more phrase and word alignments.
Studies show that a larger corpus size for training increases the quality of a Moses-based SMT system. Unfortunately large amounts of parallel training data is available for only a restricted number of language pairs making the acquisition of a large corpus difficult.

Typically, the more data is used to estimate the parameters of the translation model, the better it can approximate the true translation probabilities, which will obviously lead to a higher translation performance.

The more data we have, the longer phrases we can learn, the longer phrases we can learn the more context we can take into account, the more context we can take into account the more non compositional phrases or idioms we can handle – resulting in a better model.

http://www.aclweb.org/anthology/R11-1077

(ii) Good-quality:

Good-quality data is more statistically relevant, and bad-quality data is more statistically irrelvant. Furthermore, with good-quality learning data we can make higher quality translations, have more control over translations, vocabulary and writing style, and any issues that may arise can be more easily understood and resolved. Good quality data provides us with deep coverage of a specific domain, whereas bad quality data may have a more shallow coverage of vocabulary.

Large amounts of training data also require large computational resources. With the increasing of training data, the improvement of translation quality will become smaller and smaller. Therefore, while keeping collecting more and more parallel corpora, it is also important to seek effective ways of making better use of available parallel training data.

(iii) Representative:
If the data is representative, we have more chance of identifying idioms/colloquial language. The collected corpora are usually from very different areas. For example, the parallel corpora provided by LDC come from quite different domains, such as Hongkong laws, Hangkong Hansards and Hongkong news.This results in the problem that a translation system trained on data from a particular domain(e.g. Hongkong Hansards) will perform poorly when translating text from a different domain(e.g. news articles).

(iv)
If the corpus of small, not of good quality or not reresentitive then our statistical language models will be poor, so any any results we achieve from our models will be of poor quality also.

1(b) Study these sentences in Abma, an Austronesian language spoken in the South Pacific island of Vanuatu, and their English translations:
1. Mwamni sileng. He drinks water.
2. Nutsu mwatbo mwamni sileng. The child keeps drinking water.
3. Nutsu mwatbo mwegalgal. The child keeps crawling.
4. Mwerava Mabontare mwisib. He pulls Mabontare down.
5. Mabontare mwisib. Mabontare goes down.
6. Mweselkani tela mwesak. He carries the axe up.
7. Mwelebte sileng mwabma. He brings water.
8. Mabontare mworob mwesak. Mabontare runs up.

Assume the following additional lexical entries:
sesesrakan: teacher
mwegani: eat

Translate the following sentences into Abma:

i. The teacher carries the water down.
Nutsu sesesrakan mweselkani sileng mwisib.
ii. The child keeps eating.
Nutsu mwatbo mwegani.

1(c)
Describe in your own words how you produced these translations, focussing in particular on the particular types of inferences you made from the parallel data provided. How is this analogous to how SMT works?
anchor words -> word pairs/alignments -> reduce search space -> repeat

- Find as many patterns of a word as possible
- Align these words to each other
- Reducing search space
- Process of elimination if cant find a word alignment for a given word
- Some words may have 1 - many relationships
- Show different alignments that youve created.

Potential other questions:
1: How might you argue that "The time for MT is now"? Justify your answer by providing three use-cases where you claim that MT is the only solution, i.e. that there is no place for human intervention in the translation pipeline for such use-cases.

- EU need loads of translators by 2020 (get stat), wont be able to reach their target therefore MT needs to step in.
- Google translate processes a billion translations a day for 200m users.
- This surpasses what professional translators handle in a year therefore it is not possible for human translators to keep up with the demand of the translations needed.
- During the fifa world cup in 2014 MT systems were able to translate 2.8m words a day which is the same as 1,134 human translators working full time for 30 days.

Question 2

2(a) Why is MT important in this day and age? Provide three use cases.

http://www.computing.dcu.ie/~away/PUBS/2013/Way_ASLIB_2013.pdf

MT is a tool that enables people to have information about a variety of things in different languages. Their is a huge demand for instant translation of material, especially on the web. 52% of online material is in English, meaning that 52% of the web is essentially meaningless to non-english speakers without automatic translations.

- **Internal Communication:** translation of emails, online chat, international communication across offices/hubs, FAQs, repetitive product descriptions such as listings,
- **Website Translation:** where rapid translation of critical updates is required, as well as for gisting purposes,
- **Bids/Tenders** - translation for gisting purposes,
- **Legal/Government Documents** - e.g. EU required to have translations of all documents. They have a requirements for more and more translators, but will not reach their target by 2020. MT will have to step in.

2(b) Advantages of a customised translation engine made by a company vs google translate.

- Improves productivity
- Translate content previously not feasible due to time or cost constraints.
- Reduces time to market.
- Reduces their translation costs.
- Eg Case Studies : DuDu & Capita IT and Ford & Systran/SAIC

2(c) Provide three cases where MT is the only feasible option.

- **Automated translation of web pages** - not feasible to manually translate entire web pages every day
- **Quick translation of a small amount of text -** emails, tweets, comments on Youtube/Facebook
- **Translation of mass quantities of data where the translation is needed ASAP - (Real time translation)**

e.g. 2014 world cup, 2.8 million words a day translated. Doesn't need to be 100% accurate. Would have taken 1,134 human translators working full time for 30 days to translate the same amount.

2(d) What is the role of a human in the MT pipeline?

Humans are used in **post editing** situations where light post-editing and full post-editing is carried out to make the MT output an understandable reflection of the source. Added stylistic niceties are added in the full post-editing scenario.
Humans are also used in the **evaluation** of machine translation systems. They are given the output of a MT and their task is to assess the quality of the output.
Human evaluators are asked to evaluate:

- *Adequacy:* Does the output convey the same meaning as the input sentence? Is part of the message lost, added, or distorted?
- *Fluency:* Is the output a good fluent sentence in the target language? Involves grammatical correctness and idiomatic word choices.

2(e) Why are some people still against MT?

There are two main problems with machine translation: the problem of **ambiguity**, consisting of a word having more than one meaning, lexically ambiguous or a phrase or sentence being able to have more than one structure, **structurally ambiguous**. And problems that arise from structural and lexical differences between languages: the problem is that some languages use different structures for the same purpose, and other times, the same structure for different purposes. Obviously, a verbatim translation can´t be the correct one, it can´t solve the problem.

Another problem is that of the **idiomatic phrases**. Some of them can be quite ambiguous, and in the case of idioms, in a literal translation the meaning of this is lost (more of a cultural thing than anything!). Ex: does this ring the bell? Literal: is the bell ringing due to something? True meaning: is this familiar to you?

Professional translators are largely against the idea of MT systems as they are **fearful** that MT systems will replace their **jobs**. Humans will still be needed in order to post edit the outputs of MT systems in order to make them better. Translators are dying off may need MT systems to be able to translate old and rarely spoken languages as there may be no translators for these languages in years to come.

Question 3

Question 4

4(a) Calculate n-grams

- p(a) = number of occurances of anumber of tokens in corpus
- p(b|a) = p(a) p(b)p(a)= bigram probability of “a b”
- Sentences:
  - <s> Denis likes Ada </s>
  - <s> Ada likes Richard </s>
  - <s> Ada likes Java </s>
- Bigram LM:
  - p(Denis | <s>) = p(<s>) p(Denis)p(<s>) = (3/15)(1/15)(3/15) = 1/15

4(b) Calculate sentence probability

- p(b|a) = number of occurances of "a b"number of occurances of "a"= bigram probability
- p(c|ab) = number of occurances of "a b c"number of occurances of "a b"= trigram probability
- Sentence Probabilities:
  - p(<s> Denis likes Ada </s>)
    = p(Denis | <s>)p(likes | Denis) p(Ada | likes) p(</s> | Ada)

4(c) Calculate sentence probability with smoothing

- Notation:
  - p = p(wn | w1, …, wn-1)
  - c = count of n-gram (w1, w2, …, wn) in corpus
  - n = count of history (w1, w2, …, wn-1) in corpus
  - v = vocabulary size
- Add-One Smoothing:
  - p=c + 1n + v

- Add-alpha Smoothing:
    - More realistic than add-one counts
    - = some value <1
    - p = c + n + v

4(d) Compare a Neural Model to a standard SMT Model

- In neural network models words are represented as multidimensional vectors in a continuous space.
- Neural network trained to predict a new word (vector) given a sequence of words (vectors).
- In n-gram models words are treated independently - if we have an unseen word we use smoothing. In neural models words are treated continuously - if we have an unseen word, we can use information from related words in the continuous space.

4(e) How can language models be evaluated and how do we know if it is a good or bad model?

Perplexity language modeling

Question 5 - Piss again

5(a) Word Alignment

5(b) EM Algorithm

5(c) Word Alignment Grid

Question 6 - Piss

6(a) Given a phrase table, sentence, constraints and search space, calculate the probability of certain hypothesis.

6(b) What is the best root?

6(c) How can we combine hypothesis to make sure of termination?

6(d) What is the obvious group to pick?

6(e) Given the small number of hypotheses, which of them should get pruned?