Language modelling's generative model: is it rational?

Karen Spärck Jones
Computer Laboratory, University of Cambridge

June 2004

This note considers issues of principle that arise in applying so-called language modelling to different language and information processing tasks. Language modelling is familiar and effective for some tasks, and has been proposed and appears promising for others (see, e.g. Croft and Lafferty 2003). However attention has been more often focused on technical manipulation than on the underlying justification for taking a language modelling view of a task. The aim of this note is to go right back to what language modelling looks like as a theoretical foundation for different language and information processing tasks.

The discussion is quite informal: there is a more formal view in the Appendix.

Language modelling as she began

The language modelling (LM) approach to speech recognition has been unexpectedly successful: this simple, statistically-based strategy, when supplied with enough training data, some smart algorithms, and hefty machine power has really delivered the goods (Young and Chase 1998). But this is perhaps not so surprising since the generative account of speech production that underlies it is very plausible. People want to communicate their thoughts, which when using language necessarily delivers a linear string of words; and while the transmission of their message may be adversely affected by noise (poor speaker articulation, concurrent background conversation, partial hearer deafness), the notion that the hearer's job is to recover the original signal that generated what they receive is reasonable, because both speaker and hearer can be assumed to be to be concerned with the same word string. This is a very crude version of a more complex reality which ignores the important matter of mapping a sound sequence to a word sequence. But it is nevertheless one which captures the essential truth about speech recognition, at the right level of detail for present purposes, namely that its aim is to identify words.

In the last decade, this view of the relation between two linguistic objects has been generalised by being applied to other language and information processing (LIP) activities. Its power, for these purposes, is that it deals with the surface manifestations of language use. Thus while we can allow that LIP 'really' deals with concepts, for which words are not wholly reliable indicators, it is equally the case that words are not wholly unreliable as content indicators, and they have the advantage of being accessible and in large supply, so patterns can be sought in their distribution. Even if deeper models of language are required to *interpret*

surface objects, applying these can be treated as a separate activity involving the human user. Before engaging the user, much can be achieved for many practical tasks, simply by applying observed word patterns: as political oratory demonstrates, using the right words can have the right effects without any intervening understanding, and this has been equally effectively demonstrated in LIP by statistically-based document retrieval.

The LM approach in LIP was initially extended from speech recognition to machine translation, and has more recently been applied to document retrieval and text summarisation. This has involved rather imaginative interpretations of the key generative idea in LM. My object here is to ask whether these generative accounts of the various LIP tasks are convincing as models of the tasks and whether, if they are not altogether plausible, this matters as long as applying the LM strategy is practically effective.

As noted, the essential notion in LM is *recovery*: some signal has been received, which has been corrupted in transmission to its receiver, and the receiver has to recover the original verbal *generating* signal. As with the speech case, in all the further tasks considered here the recovery process is treated as one of recovering the original word string from the manifest word string, though it is evident that LM can be treated in a much more abstract way as implicitly recovering the arbitrarily complex concepts that underlie actual word strings, for which the words are merely convenient indicators and not Dinge an sich. Ths issue about the plausibility of the LM account is thus that of the treatment of the task as a recovery process. In particular, is this an apposite and productive metaphor or merely, as sometimes seems the case, a transfer from the intuitively appealing speech case of a technical notion which does not, in the new case, properly have an analogous intuitive interpretation?

Language modelling accounts of tasks

Machine translation

In machine translation (MT) (Brown et al. 1992, Germann et al. 2001), the given is the *source* language discourse and the aim is to replace this by the 'correct' *target* language discourse. In practice, the discourse-discourse mapping is a text-text mapping, and in what follows I shall assume for simplicity that we are only dealing with text. The recovery account of MT is that there was in fact an original text - the unknown target language text - which generated the actual source language text we now have, so what we have to do is use the latter to regain the former. The operational method is to exploit data where we already have paired source and target texts, observe what the target texts are like for the source texts and then exploit this observation data by analogy to produce the target text for a new source.

The sleight of hand in this strategy for MT is cool but also rather odd. What we have is a source text: but why should we assume that it is the crufty corrupted version of something we haven't got already? What reason do we have to regard some unknown text as the pure well of expression undefiled that has alas, by the time it reaches us, been muddied? The LM account is at variance with the normal reality which is where, for example, we have our Shakespearean text and struggle to render it into Turkish. We do not believe that Shakespeare really wrote in Turkish and all we have is some faint malformation of his original?

The LM view of translation can can only be taken at a very abstract level, which does not refer to any actual or hypothesised prior text. But then what is the model actually

assuming? All it can be assuming is that there could be some text (the target text) which expresses some ideas more or less adequately, that the text we have (the source text) also expresses these ideas more or less adequately, and that the best fit we get by constructing the target from the source is a more or less adequate expression of the ideas in question. There is of course no presumption that either text is the best possible expression of the ideas, and also no guarantee that the target is as adequate as the source. The text-to -text mapping can only be justified on this basis even if the operations involved make no reference to concepts, only surface strings.

But this account of MT is a perfectly conventional one. It is hard to see it as an LM communication one if the original generating text is in fact wholly imaginary. It would be more in tune with the *general* LM way of looking at things to say that the original is a conceptual message to be recovered through a transparent verbal expression, but this strategy precisely throws away the the simplicity of dealing *only* with text-to-text relationships. Thus one might argue that while implementing LM as a way of doing MT as described earlier may be very successful in practice, as an intellectual account of MT it is all hocus-pocus.

In both the speech recognition and translation cases there is rough parity of length between the two texts, and also rough equivalence between local units (words, phrases, sentences). With the other two tasks, document retrieval and text summarising, this situation is quite different, with one member of the pair much shorter and less complex than the other. This reflects task realities and requirements, but raises further questions about the justification for the LM generative model.

Document retrieval

The LM view of document, or text, retrieval is that the document to be preferentially retrieved is that which is most likely to have generated the user's query (Ponte and Croft 1998, Berger and Lafferty 1999, Miller et al. 1999, Hiemstra 2000). (I am assuming only one relevant document for simplicity, not as a matter of theory.) The notion is that the user has the whole document they want in mind, but their initial version of it has been sadly corrupted to become the poor little thing that a user request normally is. The function of retrieval is to recover the whole pristine original document.

There is no reason in principle why the retrieved text should be substantially longer than the user's query, and indeed a whole text might be submitted as a query on a 'more like this' basis; but in practice, with good cause, queries are normally much shorter and simpler than returned documents. Thus retrieval in the LM view is recovering the 'real' full document from the very reduced and partial version represented by the query. By comparison with the recognition and translation cases, the original signal has been much degraded. This may not only be because the query is short but because it consists only of a set of terms, either as submitted by the user or where, as is usual for sound reasons in retrieval systems, the document and query texts are replaced by bag of words surrogates, i.e. simple index descriptions without discourse syntax.

The problem with the LM idealisation of retrieval as recovering a whole document from its very minimal received representative is illustrated by comparison with the speech case. Suppose that a speaker has uttered an extended discourse, and the transmission channel is so noisy that the hearer can identify only a few scattered words, some different, some repeats,

with any feeling of confidence that these words at least were actually uttered. This could be because they were emphatically spoken, but also because they could all be seen as associated with a common domain of discourse. However unless the hearer has strong independent leads, say from the surrounding particular physical context, in going beyond the odd words and hypothesising the speaker's message, he does not have an adequate base, in a few keys, for reconstructing a whole 'original' document in all its discourse detail. I'd defy anyone, even when mentally supplementing "wealth" and "nations" with some notions about this having to do with property, trade and so forth, to actually have the whole of Adam Smith's classic treatise in mind.

Thus in the speech case, the hearer normally gets far more than the odd word, even if they don't know in advance what the speaker actually said. In the retrieval case, the user is also ignorant, but the further presumption is that the user is in a much more awkward pit of of ignorance, suffering from a serious information gap or anomalous state of knowledge. There is therefore a very strong reason to argue that the LM account cannot amount to more than saying that the user can *recognise* that a document, when he reads it, meets his information need because it has supplied the information he did not have before he read it. But it is hard to square this with the notion that the document generated the query. There is something distinctly peculiar about the strictly generative account, because if the user really had a substantive internal version of the document already, he would not need to do retrieval at all. Of course a user might want to return to a document they had already seen, as in known-item searching, but this is not what retrieval systems to find new relevant documents are about.

Categorisation, filtering (routing) and tracking

These might appear, as tasks, to be variants on the retrieval task, as in Teahan and Harper 2003's view of categorisation, for example, and thus not to raise any new issues at the 'deep modelling' level.

Categorisation treats each class separately, and while yes/no assignment differs from retrieved document ranking, the essential relationship between class characterisation and document is like that between retrieval query and document, differing only in (usually) being more substantial on the class characterisation side.

Filtering can be viewed simply as a temporal succession of (de facto closely related) requests, with the system deciding whether a new incoming document generated the current query or not. The user then, after assessing this document, proceeds to or stimulates a new query: feedback modification of the query over time is perfectly natural.

Fully dynamic filtering can also be viewed as topic tracking, since the user's query modification is a response to the actual changes in incoming literature content. But topic tracking, as illustrated by the recent Topic Detection and Tracking (TDT) evaluation for news stories, and to which LM has been applied (Kraaij and Spitters 2003), raises interesting questions for the LM approach.

There is nothing obviously problematic about news story *tracking*. Kraaij and Spitters take a given text story as their starting point, but the query analogue need not be a conventional text. It could be, say, a Schankian-style structured event scenario, just as a retrieval query may be a Boolean expression rather than a topic text. We can thus imagine tracking to reflect a succession of instantiated events and subevents so we are increasingly well placed to predict,

i.e. recognise, the story continuation.

This seems plausible enough. But now consider the larger and more complete form of the task, namely *detecting* a new story, not just tracking one to which we have been alerted. It is not at all clear how this can be handled within the LM framework. What does the user have in mind *before* the first incoming item about a new story? How is this new story recognised as such on the LM account? We might have some expectations about news stories as a genre, i.e. that they will be about, say, political scandals or rock music celebrities or ... But it is not obvious that this is very usefully or properly expressed in the LM 'recognition' paradigm: it would seem to imply that I have expressed my topic as a large disjunction of all the types of things news stories can be about, but excluding those that are already running this week, and I recognise a new story as new because it pulls out one of the other disjuncts. Or in the generative view, the disjunct representing a particular actual new story was so messed up by the generation process that all that reached the other end was the set of possible sorts of stories. Or maybe nothing got through the generation process and all I start with is a tabula rasa. But all of this is very peculiar.

(There are also interesting questions about the training data for a new story recognition system: how much data would be required to establish reliably that if X has been talked about for some time, Y (= anything but X) is going to appear?)

A rather similar situation arises when relevance assessment for ordinary retrieval is further constrained by a specific requirement for novelty. Even allowing for some restriction of novelty in this case to 'novelty within the broad subject area already established by previous relevant documents', the space of things that could be novel is pretty wide open.

Question answering

This is currently being treated, for evaluation, as a upstanding system task in its own right, rather than as a necessary LIP system subfunction, and LM has already been taken as an appropriate strategy for doing it (e.g. Croft 2003). While some sorts of question answering might be assimilated to summarising (see below), for example in responding to a 'Who was X' or 'What is Y' with a descriptive paragraph, question answering may also be viewed as a tighter form of retrieval.

But pursuing this line makes the assumption on which LM is based stronger and more problematic. If answering a question is treated as recovering what the questioner already knows, why did he ask the question? [1]

Text summarisation

The LM version of summarising, with one text much shorter than the other, is the reverse of the retrieval one. Here the idea is that the generator, to be recovered, is the crisp, succinct expression of the crucial idea underlying a text. This core message is then messed up in the full text composition by being extended with detail - elaboration, repetition etc. (Banko et al. 2000, Knight and Marcu 2000, Mittal and Witbrock 2003). Thus the objective in summarising the given extended text is to cut out all the waffle and recover the pure core

[1] This is not the place to consider LM as a psychological memory model.

original.

This account of summarising is very well illustrated by *The Guardian* newspaper's jokey summary of the week's fashionable novel in its 'Editor' supplement. But the joke in part depends on seeing how a quite long initial description of the novel can be encapsulated in a five-word one: given only the five word capsule, the full novel could be about pretty well anything. Certainly the suggestion that a serious author does nothing more than pad out a plot line seems rather thin: is there nothing more to Henry James' *The Golden Bowl* than 'dubious marriages in high society'? Similarly, most of us would not accept that there is no more required of a scientific paper on retrieval than, say, 'language modelling works'.

In fact in the LM account of summarising, even more than in the retrieval case, there is no characterisation of the precise nature of the relation between the two texts: they are just generator and generatee. But the longer text is not just some arbitrary enlargement of the initial brief one (or in the conventional familiar direction, the shorter one some arbitrary abbreviation of the longer one). The presumption is that the brief one is the clue, or cue, to the longer one, by embodying its core notion as expressed in some particular words. The longer text expands rather than merely addles the shorter one. Just as with retrieval and translation (and also speech recognition when viewed as transcription), the crucial characterisation of the relationship between the input and the output is in fact offloaded in the LM approach onto the choice of training data. We can use LM for summarising becase we know that some set of training data consists of full texts paired with their summaries. When some instantiation of the generic model has been created by training on some body of data, we can say that this does embody the particular character of the summarising relationship: the training data gives us a realisation of the notion of summary text as core of the full one. But without the data LM tells us nothing about the task.

It is thus not clear how LM gives us insight into the nature of the individual LIP task, even if it allows us to replicate the results of humans doing it.

Assessment

Language modelling as deep generation

It is not difficult, given its observed effectiveness for some tasks, ever more training data, and developing software, to see how the language modelling treatment might be further pursued in practice for the tasks just considered; or, given the way it has spread so far, how it might be proposed for other LIP tasks.

There has been some controversy about the propriety of LM as a task model, but this has focused on issues in characterising specific tasks - for instance on the primitive status of relevance in the retrieval case, or on the reality of multiple relevant documents for the same query (see Lafferty and Zhai, 2003, Sparck Jones et al. 2003). My concern here is with LM as a *generic* LIP task model: from one point of view LM is an abstract, formal apparatus, but using it successfully implies modelling legitimacy. This, for LIP, relies on a common notion of text (or quasi-text) object pairs, which has in turn been taken to presuppose a common generative-recovery relationship between the members of a pair.

My task analyses suggest there are serious problems about this 'deep generation' framework. The robust response is to say bother modelling language modelling, and take the line

that LM is just some convenient technology which can be applied, after training, to produce new texts of some sort in place of other texts of some other sort. So all that is needed to justify LM, for as many tasks as possible, is that it can be implemented with some practical success. From this point of view, if LM works well in task evaluation, it is sufficient that LM is a good way of exploiting training data to estimate probabilities. Once the step has been taken to treat task processes as operations on surface statistical data, getting probabilities right is what counts.

Using surface data might be seen as ruthless pragmatism, where explanatory models are lacking. But where LM is successful, as it already has been for some tasks, this seems to imply that there is some LM-grounding model for the task. This has already been a matter for discussion in retrieval, where there are competing model-based technologies (Robertson and Sparck Jones 2004). It might turn out that LM technology only works well for 'rough' tasks like retrieval or simple task forms like 'keyphrase' summarising, or for limited task applications like highly constrained dialogue inquiry (Young 2002). But even these cases seem to imply that LM captures something about LIP tasks, if not everything.

So if the generation-recovery account of LM is unconvincing, is there a better one?

Language modelling as paraphrase

The tasks I have considered relate pairs of texts (or quasi-texts) under some general meaning preservation constraint, so they can be seen as variations on the generic notion of *paraphrase*, each with its own particular twist. Normally interpreted, paraphrasing has to produce something different from but as good, in some required way, as the original text. Recasting my initial account of LM in paraphrase terms means that we are stuck with what we cannot always suppose is a good paraphrase, and have to figure out the no less satisfactory, or superior, original. We may in principle be dealing with alternative representations of the same non-linguistic concepts, this is neither here nor there, because we have no direct access to the concepts and so cannot, strictly, use them to guide the paraphrase process. We may have is some adumbrated output procedures that could have given us what we finally have, not the initial text they had to work on; and we may, further, have some general characterisation of the kind of transformation - reductive or expansive, say - that motivated the paraphrase we have.

But why should we preserve the directionality of the original LM account, dressed up as paraphrase? It doesn't get us anywhere in making the generative view of LM more plausible as a common form of task model.

It seems much more sensible to abandon the recovery view of LM and talk, instead, and talk about producing a new, derived text from some given text, i.e. apply generation in the normal sense implemented in e.g. translation or summarising programs. We don't recover a lost summary: we make one.

Paraphrase looks much more plausible, on the face of it, this way round. We have some text and derive another according to task requirements, elaboration in retrieval, reduction in summarising, gap-filling in question answering, etc. Translation gives a version of the original in new languages, and speech recognition can also be treated as a language transformation. The space of possible paraphrases for some original will be very large, but with suitable task guidance we can hope to get something as output that is good enough without having to hunt for the ideal, i.e. one correct, paraphrase. Explicit task guidance may be hard to come by, but this does not matter because existing exemplars (in large quantity) are a probably

superior substitute.

But even when driven in a more rational direction, treating all the tasks are paraphrastic text transformations is not very appealing. It savours too much of 'lit crit' structuralism, taking the initial text at its surface value and playing games with it. Moreover, as all the tasks are defined by their own distinctive and in some cases very different transformational requirements, calling it all paraphrasing does nothing to explain why LM is a good common explanatory model for LIP tasks as opposed to a way of manipulating texts by exploiting observed relations between pairs of texts, however good LM may be at this. For example, we have already to know what summarising is, in order to legitimise the training data for LM-based summarising.

Language modelling as expression seeking

One possible way forward is to bite the bullet that text expresses concepts, and talk of a *'struggle for expression'*. The user has some notion of what they are seeking as a representation of some concepts, because they have some knowledge of some representation language and some appreciation of what they would like their representation to achieve. However the user's grasp of what they are about may be more or less adequate, depending not only on their understanding of the representation language and the task requirements, but also on what they think the concepts they are dealing with in a particular situation are: there is plenty of scope for muddled ideas and inadequate expression. For instance, in the case of retrieval or question answering, the user tries to sketch or frame their current holey concept set. In the case of translation they utter in English - as it were slowly and clearly, like a hopeful tourist, - or maybe Franglais, as their best approximation to French.

This approach to LM has been put to me, more substantively and properly, by Rich Schwartz. [2]. It was also briefly invoked as providing a basis for LM in Ponte and Croft (1998)

Thus if we consider document retrieval, the basic idea is that while the user does not know actual relevant documents in advance they have an idea of them in vocabulary terms: that is, the user has some model of the *distribution* of terms in documents that would be relevant to their information need, and applies this model in formulating their query. However as the user is guessing, their query may include terms that are not in all relevant documents, terms that are not in any relevant document, and so forth. We can say that the user's search terms are drawn partly from what may be called the 'relevance language' distribution and partly from what may be called the 'query language' distribution where, in the absence of better data, we may guess that the latter looks something like the corpus language distribution.

Thus not only is the user's vocabulary distribution model at best an approximation to the actual relevance language distribution model; they may also differ systematically. Characterising the user's model in detail, and leveraging a better relationship to the actual relevance language model, then leads to ways of refining the distributional picture, e.g. by separating interesting (aka content) words from uninteresting (aka stop) words, dealing with variant words forms or synonyms, and so forth. Any specific data, e.g. about the actual occurrence of terms in relevant documents, or about the query language drawn from a large query sample (as in Scholer and Williams 2002), can be used to help this refinement.

This 'deep distributional' model can also be made more subtle by thinking of documents and queries as dealing with subjects, and of the document and query terms as drawing on

[2]personal communication December 2002

subject vocabularies, so the user is thinking not just of a distribution of terms in (relevant) documents but of terms for subjects that (relevant) documents are about. There will thus be two distributional models reflecting the document and query versions of the subject vocabulary (at least: there could be more complicated situations with e.g. multi-subject documents).

A similar account can be given of summarising: there is a notion of a stock of information which is more or less fully developed in different versions, longer document or shorter summary (and one can also have longer or shorter summaries).

The idea of characterising tasks as attempting to capture language distributions over notionally common content is certainly more attractive than the rude recovery accounts; and it allows in principle for language of any complexity, i.e. not just simple word unigrams but higher order units and multigrams. The use of LM serves to focus attention on the range of distributional models that are implicit in tasks and their data and, using whatever data is available, makes it possible to refine these models, and to relate them in order to carry out the task.

So far, language models for tasks have been, even in their most refined forms, very crude. But they are a useful starting point, and thinking of the underlying generic situation as one of attempting to choose appropriately distributed words (or, more generally, linguistic units and sequences), when engaged with a task, is a rational underpinning for technology that has to work with how words (units etc) are actually distributed in whatever data may be at hand to support or train for the task.

Conclusion

This generic underpinning for LM is certainly more attractive than the recovery one. When we look carefully at what the deep generative account of LIP tasks is claiming about them, that model fails to fit. It is not clear, on the other hand, what the expression-seeking account really achieves in *explaining* whatever the various tasks have in common as language-using activities. It may rather serve better, through drawing attention to the sets of language distributions involved, to promote careful analysis of what is going on in particular tasks.

From that point of view, pushing backwards from the surface data can be helpful whatever treasure may be buried in the deep rich earth of foundational models, if only we could find it. So we may just choose to adopt LM as a technology to apply to get one text from another. It is certainly being implemented in this style, in ever new directions, in ever more variations, and with instructive success as well as engaging promise (e.g. Barzilay and Lee 2004). There is also no reason why LM should not range over more complex objects and orderings than, say, word bigrams, as has already been recognised for translation and summarising (e.g. Knight and Marcu 2000). This should make LM more powerful, though it may not be easy to do in practice.

As a technology, LM appears to make very good use of whatever training data can be got, and as ever more data appears, this is to LM's advantage. But it is also very robust and can do without any training data in the strict sense.

Even so, nothing that has been done so far in using LM for LIP tasks, even of limited kinds, shows that LM fully captures all there may be to the task. So it is likely to reach a performance plateau. This may be quite high and may give the user quite a nice view of the surrounding landscape. But it will be somewhere on the side of the mountain, not on the top.

We can't expect LM to do any better because, regardless of how one tries to rationalise LM, it has the crucial limitation that it depends on data volume and is essentially dealing with averages. That means there is no reason to suppose it will match any individual situation really well.

Acknowledgement

I am very grateful to Steve Young, Steve Robertson and Rich Schwartz for their comments and suggestions.

Appendix

In the Note I considered the model(s) underlying language modelling quite informally. The key points can also be made more formally by referring to the dependencies in the probability equations involved.

Thus one way of setting up the statistical framework for LM follows from the information/communication theory view and thinking about channel models. This is classical in the speech recognition case, and when applied to MT quickly leads to the formulation

$$\hat{T} = \underset{T}{\text{argmax}}\{P(T|S)\} = \underset{T}{\text{argmax}}\{P(S,T)\} = \underset{T}{\text{argmax}}\{P(S|T)P(T)\} \tag{1}$$

where S is the source and T is the target language. $P(S|T)$ is a sort of likelihood function which is a measure of how well the observed data (the source) matches the proposed target text T. $P(T)$ is the so-called language model term which says how well the proposed target text fits the known distribution of target texts. The likelihood function is typically constructed relative to an alignment of the two texts decomposing it into a product of factors which embody the probability of various alignment effects such as displacement, inversion etc and translation between corresponding lexical items of the two texts. $P(T)$ is trained on a large corpus of material in the language T and its inclusion insures that all hypothesised T's are reasonably well-formed in the language.

In graphical terms, the simple channel model corresponds to the intuitive-looking picture of fig 1.

Figure 1: Comms Channel View

However the transformations of eq 1 suggest that the dependence is the other way round, from T to S. But while the resulting directional interpretation is natural for speech it is, as the Note argues, much less convincing elsewhere. Thus for the MT, the presumption that there is a text T fitting the current new S like past Ss and Ts fitted one another, though superficially plausible, presents problems on deeper analysis; and the analogous account is even less plausible and more questionable for other tasks, though they have been presented as other applications of a *generic* 'translation model' for text pairs (as in Berger and Lafferty 1999, Banko et al. 2000).

The alternative statistical framework is the Bayesian one. Here the observations result from sampling conditional distributions and the conditioning variables need not, and indeed often are not, observable. Thus, the Bayesian approach to MT might be to assume that any sentence S in the source language and T in the target language derive from a single underlying meaning M. So we have $P(S|M)$ and $P(T|M)$, but T is conditionally independent of S. In graphical model terms this gives the picture shown in fig 2.

In this case, given only S, we can only hope to find the joint distribution of T and M ie $P(T, M|S)$. But M is not known and we don't want to know it, rather, we want $P(T|S)$. So like all good Bayesians when we don't know something, we simply integrate over all possible values of M ie

$$P(T|S) = \sum_{M} P(T, M|S) \tag{2}$$

Figure 2: Bayesian View of MT

Now we can find T!

$$\hat{T} = \underset{T}{\text{argmax}}\{\sum_{M} P(T, M|S)\} = \underset{T}{\text{argmax}}\{\sum_{M} P(T|M, S)P(M|S)\} \tag{3}$$

and since T is conditionally independent of S when M is known this becomes

$$\hat{T} = \underset{T}{\text{argmax}}\{\sum_{M} P(T|M)P(M|S)\} \tag{4}$$

Of course, we don't know $P(T|M)$ or $P(M|S)$ so this formulation has no practical interest, nevertheless, the equation looks intuitively pleasing – T depends on M and M depends on S, we do not know M so we sum over all possible values and the net result is the probability $P(T|S)$ that we were looking for.

But what can we do in practice to build a translation model? Well, eq 3 can be rewritten as

$$\hat{T} = \underset{T}{\text{argmax}}\{\sum_{M} P(S|T, M)P(T|M)P(M)\} \tag{5}$$

If we now say (incorrectly) that $P(S|T, M) = P(S|T)$, then eq 5 becomes

$$\hat{T} = \underset{T}{\text{argmax}}\{P(S|T)P(T)\} \tag{6}$$

which is the channel model. Our assumption $P(S|T, M) = P(S|T)$ is equivalent to saying that the mutual information between S and M given T is zero.

Does the Bayesian model do anymore than just add some insight into the approximation incurred by the channel model? Well, potentially it does. We don't know $P(X|M)$ but we can invent one. It could be a hidden Markov model, or a hidden Hierarchical model, etc. If we had enough data, we might train something (using EM) which does better than the channel model. The key point here is that we don't need to observe M directly in order to learn and use an approximation for $P(X|M)$ because whenever we need to use this function we integrate over all possible M.

On the other hand, if we get the model wrong (and this is probably easier than getting it right) then the results will be worse (cf the difficulty of designing a hierarchical model to outperform a simple n-gram in speech recognition). This is the reason for jumping straight into the channel model as a first approximation. It's the model of ignorance and as in life, it is better to know what you don't know than it is to think you know something that you don't!

References

M. Banko, V. Mittal, and M. Witbrock, 'Headline generation based on statistical translation', *ACL 2000: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 2000, 318-325.

R. Barzilay and L. Lee, 'Catching the drift: probabilistic content models, with applications to generation and summarisation', *HLT-NAACL 2000: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 2004, 113-120.

A. Berger and J. Lafferty, 'Information retrieval as statistical translation', *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*, 1999, 222-229.

P.F. Brown et al. 'Class-based n-gram models of natural language', *Computational Linguistics*, 18(4), 1992, 467-680.

W.B. Croft, Presentation at Advanced Question Answering for Intelligence (AQUAINT) 24-Month Workshop, December 2003.

W.B. Croft and J. Lafferty (Eds.), *Language modelling for information retrieval*, Dordrecht: Kluwer, 2003.

U. Germann et al. 'Fast decoding and optimal decoding for machine translation", *Proceedings of the 39th Annual Conference of the Association for Computational Lingustics (ACL 2001)*, 2001, 228-235.

D. Hiemstra, *Using language models for information retrieval*, PhD Thesis, Centre for Telematics and Informatio Technology, University of Twente, The Nethrelands, 2001.

K. Knight and D. Marcu, 'Statistics-based summarisation - Step 1: sentence compression', *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI 2000)*, 2000, 703-710.

W. Kraaij and M. Spitters, 'Language models for topic tracking', in Croft and Lafferty 2003.

J. Lafferty and C. Zhai, 'Probabilistic relevance models based on document and query generation', in Croft and Lafferty 2003.

D.R.H. Miller, T. Leek and R.M. Schwartz, 'A hidden Markov model retrieval system', *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*, 1999, 214-221.

V.O. Mittal and M.J. Witbrock, 'Language modelling experiments in non-extractive summarisation', in Croft and Lafferty 2003.

J.M. Ponte and W.B. Croft, 'A language modelling approach to information retrieval', *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, 1998, 275-281.

S.E. Robertson and K. Sparck Jones, 'Retrieval system models: what's new?', in *Computer systems: theory, technology, and applications*, (Ed. A. Herbert and K. Sparck Jones), New York: Springer, 2004.

Scholer, F. and Williams, H.E. 'Query association for effective retrieval;, Proceedings of the Conference on Information and Knowledge Management (CIKM 2002), 2002, 324-331.

K. Sparck Jones et al. 'Language modelling and relevance', in Croft and Lafferty 2003.

W.J. Teahan and D.J. Harper, 'Using compression-based language models for text categorisation', in Croft and Lafferty 2003.

S.J. Young, 'Talking to machines (statistically speaking)', ICSLP 2002.

S.J. Young and L.L. Chase, 'Speech recognition evaluation: a review of the U.S. CSR and LVCSR programmes', *Computer Speech and Language*, 12, 1998, 263-279.