A Hierarchical Language Model for CSR

Francisco J. Valverde-Albacete[1], *José M. Pardo*[2]
[1]Dpto. Tecnologías de las Comunicaciones, Univ. Carlos III de Madrid, Spain
[2]Grupo de Tecnología del Habla. Dpto. Ingeniería Electrónica, UPM, Madrid, Spain

ABSTRACT

We present a new language model that includes some of the most promising techniques for overcoming linguistic inadequacy, - including POS tagging [3] and refining [4], hierarchical, locally conditioned grammars [5], parallel modelling of acoustic and linguistic domains [6] – and some of our own: language modelling as language parsing, and a better integration of the whole process with the acoustic model resulting in a richer educt from the language modelling process.

We are building this model for a translation into Spanish of the DARPA RM task, maintaining the same 1k words vocabulary and some 1000 sentences.

1. INTRODUCTION

For the purpose of traditional ASR, a 'language model' is just the expression of collocation constraints in terms of the language structure, but in our opinion it must account for the expression of subclassing and structural constraints as well. The result of the language model, a *candidate sign* or *hypothesis*, should not only be a sequence of units but a whole linguistic analysis.

However, although highly unsatisfactory from a linguistic point of view, the n-gram model has challenged all attempts to improve on its linguistic modelling capabilities [1], possibly with the exception of decision trees [2].

Our belief is that one has to complicate the language structure being modelled to improve the quality of the language model. This new, more complex structure should, however, profit by the well-known advantages of less structured models such as n-grams that capture well constraint locality.

We are building this model for a translation into Spanish of the DARPA RM task, maintaining the same 1k words vocabulary and some 1000 sentences. The reason why we are using such a reduced task is that the data needed for this model is a very costly treebank in the sense that the analysis have to be made by hand and the POS tags must be carefully designed to comply with the ordering constraints to ensure the correct calculation of lower level signs.

In this paper we first propose a language structure and modify accordingly the problem of language modelling; then we propose a probability decomposition for the problem and sketch future work.

2. THE MODIFIED PROBLEM

1. The language structure

Linguistic entities – meanings, significants - show a tendency to be related to units of the same complexity level – either structural or functional -, and the sets of entitities thus related can be considered a *hierarchy of levels*, a level being a set of units of similar complexity or functionality. For the purpose of the application at hand we only distinguish the following lexical (semantic) levels: *lexeme*, *phrase*, *clause* and *sentence* in ascending order of complexity; accordingly, we only distinguish the following levels of "phonetic" substance: *phonemes*, *words*, and *sequences of words*. It is important to note that any meaning unit in a semantic level may have a related significant unit in any significant level, and vice-versa; in particular "words" will not solely be linked to "lexemes", as in most language models.

On one hand, linguistic entities show two behaviours: a *paradigmatical behaviour* when opposed to units that can fulfill the same syntactic behaviour – essentially a class membership – and a *syntagmatical, or syntactic behaviour* when collocated with units of any paradigmatical behaviour – essentially a sintactic function-. The recursive nature of these two behaviours in the level hierarchy is what makes the definition of adequate classes and grammars so challenging a task.

While doing any linguistic processing a unit shows its paradigmatical behaviour in the *paradigmatical phase* of the processing; similarly the syntagmatical behaviour is only manifest in the *syntagmatical phase*. Any linguistic processing – including parsing or generation must alternate between these two phases.

On the other hand, linguistic components cast themselves into:

- Lexicalized (pre-built) hypotheses retrieved from a database of signs, ${}^{i}s \in {}^{i}L$ and ${}^{i}\varsigma \in {}^{i}\Lambda$ or
- Hypotheses built on-the-fly according to a grammar, ${}^{i}s \in \mathcal{I}({}^{i}\Sigma)$ and ${}^{i}\varsigma \in \mathcal{I}({}^{i}S)$

where $\mathcal{I}({}^{i-1}\Sigma)$ and $\mathcal{I}({}^{i-1}S)$ denote, respectively, the set of paradigmatical components built out syntagmatical components one level less complex, and the set of syntagmatical components obtained from paradigmatical components in this level. This recursive definition ends at either lowest level lexeme – for meanings -, and – phoneme, for significants.

Thus in fact the set of paradigmatical components for each level i, ${}^{i}S$, is built out of two distinct, disjoint subsets:

(1) ${}^{i}S = {}^{i}L \cup \Im({}^{i-1}\Sigma) \quad \varnothing = {}^{i}L \cap \Im({}^{i-1}\Sigma)$

and the set of syntagmatical components for each level *i*, ${}^{i}\Sigma$, is built out of two other disjoint subsets:

(2) ${}^{i}\Sigma = {}^{i}\Lambda \cup \Im({}^{i}S) \quad \varnothing = {}^{i}\Lambda \cap \Im({}^{i}S)$

Furthermore, we consider all signs to be in fact products of two component domains: the acoustic and the meaning components, i.e.:

(3) ${}^{i}s = ({}^{i}s.p, {}^{i}s.m)$

where ".p" and ".m" denote each the acoustic part and the meaning part of each candidate.

To describe the process of obtaining such analyses let us define the following objects whose exact meaning will be made clear along this paper (note that all paradigmatical variables have latin face and all syntagmatical variables have greek face):

- ${}^{i}c_j$ restriction state of candidates in the paradigmatical phase in level *i* at step *j*;
- ${}^{i}\kappa_k$ the construction state of a hypothesis in the syntagmatical phase in level *i*
- ${}^{i}s$ a candidate for recognition or modelling in level *i* – a paradigmatical sign.
- ${}^{i}\varsigma$, a unit to continue exploring a grammar in level *i* – a syntagmatical sign.
- ${}^{i}x_j$ the paradigmatical exploration state of observations for level *i* at restriction step *j*.
- ${}^{i}\chi_k$, the syntagmatical exploration state of observations for level *i* at construction step *k*.

2. The paradigmatical constraint order

We model (lexical) paradigmatical meanings with refined POS tags in the style of [4], although our tags are handcoded. By allowing refinement we are implicitly allowing an order for constraints. To further understand it, let $(\_.c: S \to C)$ denote a meaning (phonetic) projection of the signs onto constraints, and

(4) ${}^{i}s({}^{i}c_j) = \{{}^{i}s \in {}^{i}S \mid {}^{i}s.c \geq {}^{i}c_j\}$

be the subset associated to constraint ${}^{i}c_j \in {}^{i}C$ in ${}^{i}S$, and

(5) ${}^{i}s \equiv {}^{i}c_j = \{{}^{i}s \in {}^{i}S \mid {}^{i}s.c = {}^{i}c_j\}$

(6) ${}^{i}s > {}^{i}c_j = \{{}^{i}s \in {}^{i}S \mid {}^{i}s.c > {}^{i}c_j\}$

be the subsets of the sign set ${}^{i}S$ that include all signs whose meaning (phonetic) part is respectively equal (5) or strictly less (6) than a particular meaning (phonetic) constraint; clearly they define a partition on (4). Let further ${}^{i}c_0$ be the least of constraints, so that ${}^{i}s({}^{i}c_0) = {}^{i}S$.

We can obtain more restricted constraints by means of an operator that restricts the constraints in the sense that the associated set is smaller. Clearly, the set of constraints that dominate a given one form a covering of the subset of strictly less constrained signs:

(7) ${}^{i}s > {}^{i}c_j = \Sigma_{{}^{i}c_k \geq {}^{i}c_j} {}^{i}S({}^{i}c_k)$

But if we further allow the order to be arborescent as that of the strings in an alphabet, the covering is a partition as well.We also claim that paradigmatical acoustical constraints can be modelled in this way.

3. The exploration orders

On our model exploration of hypothesis is integrated with hypothesis creation (thus we reduce recognition to integrated exploration and hypothesis construction).

Let us suppose we have an order on paradigmatical exploration states in which ${}^{i}x \geq {}^{i}x_0$ that is to say, the starting exploration state is the least one from which all others stem. Then, each candidate hypothesis will only be valid in case the exploration of (acoustic) observations reaches a valid state ${}^{i}x$ from some given starting state ${}^{i}x_o$ common to all hypotheses, so that:

The syntagmatical exploration order is much more familiar: let ${}^{i}\chi$ denote a segmentation of the observation; then ${}^{i}\chi_j \geq {}^{i}\chi_k$ is a more restricted segmentation if χ_j has at least all the segmentation states of ${}^{i}\chi_k$ and possibly more.

4. The modified language modelling problem

The modified language model must therefore provide probabilities and structure for the sentences in the task. It should start therefore analyzing at a *top* level from which to recover all the rest:

(8) $p({}^{top}s \mid {}^{top}o) = \Sigma_{{}^{top}x_0} p({}^{top}x_0 \mid {}^{top}o, {}^{top}c_0) \cdot$

$\cdot \Sigma_{{}^{top}x \geq {}^{top}x_0} p({}^{top}x {}^{top}s \mid {}^{top}x_0, {}^{top}c_0)$

But if we consider a common origin for all hypothesis, as usual only one term should be used to compare hypotheses:

(9) $p({}^{top}s \mid {}^{top}o) \propto p({}^{top}x, {}^{top}s \mid {}^{top}x_0, {}^{top}c_0)$

3. INTEGRATED EXPLORATION AND CONSTRUCTION

5. Solving the paradigmatical problem

We start with a version of (9) generalized in the level and the constraint to be solved; as stated by (5) and (6), only two types of solutions are valid with respect to the constraint:

(10) $p({}^{i}x, {}^{i}s \mid {}^{i}x_0, {}^{i}c_j) =$

$p({}^{i}x, {}^{i}s \equiv {}^{i}c_j \mid {}^{i}x_0, {}^{i}c_j) + p({}^{i}x, {}^{i}s > {}^{i}c_j \mid {}^{i}x_0, {}^{i}c_j)$

First, paradigmatical units can only be further restricted provided there exist more restricted elements ${}^{i}c_k \geq {}^{i}c_j$;

therefore, if we design the order such that these subrestrictions partition the hypotheses as in (7) we have a base case for recursion in the paradigmatical phase:

$$(11)\quad p({}^i x, {}^i s > {}^i c_j \mid {}^i x_0, {}^i c_j) = \sum_{{}^i c_k \geq {}^i c_j} p({}^i x, {}^i s \mid {}^i x_0, {}^i c_k)$$

These conditions are met if the restrictions form a branching or tree partial order whose root is ${}^i c_0$. The tagset of our treebank was handcoded to comply with this tree structure, based on standard linguistic theory for Spanish.

Second, each of the two sets may have a non-void intersection with the sets of lexicalized or built units for this level as stated in (1) the net effect of which is to further partition:

$$(12)\quad p({}^i x, {}^i s \equiv {}^i c_j \mid {}^i x_0, {}^i c_j) =$$
$$p({}^i x, {}^i s \equiv {}^i c_j, {}^i s \in {}^i L \mid {}^i x_0, {}^i c_j) +$$
$$p({}^i x, {}^i s \equiv {}^i c_j, {}^i s \in \Im({}^i \Sigma) \mid {}^i x_0, {}^i c_j)$$

Third, lexicalized hypotheses for a particular restriction only have to be evaluated for the probability of occurrence of the product event $({}^i s.p \equiv {}^i c_j.p, {}^i s.m \equiv {}^i c_j.m)$ for each hypothesis:

$$(13)\quad p({}^i x, {}^i s \equiv {}^i c_j, {}^i s \in {}^i L \mid {}^i x_0, {}^i c_j) =$$
$$p({}^i x, {}^i s.p \equiv {}^i c_j.p, {}^i s.m \equiv {}^i c_j.m, {}^i s \in {}^i L \mid {}^i x_0, {}^i c_j) \approx$$
$$p({}^i s.p \equiv {}^i c_j.p, {}^i s.m \equiv {}^i c_j.m, {}^i s \in {}^i L \mid {}^i c_j) \cdot$$
$$\cdot p({}^i x \mid {}^i s.p \equiv {}^i c_j.p, {}^i s.m \equiv {}^i c_j.m, {}^i s \in {}^i L, {}^i x_0, {}^i c_j)$$

These instances of signs and their probabilities are learnt from the treebank, and no attempt is done to smooth them.

Finally, if we accept that $\Im({}^i \Sigma)$ is a language built out of the symbols in ${}^i \Sigma$, we may then comprehend the problem posed by as that of finding the strings belonging to that language to be further specified below:

$$(14)\quad p({}^i x, {}^i s \equiv {}^i c_j, {}^i s \in \Im({}^i \Sigma) \mid {}^i x_0, {}^i c_j) \approx$$

$$(15)\quad p({}^i x, {}^i s \mid {}^i x_0, {}^i c_j, {}^i s \in \Im({}^i \Sigma)) \cdot p({}^i s \in \Im({}^i \Sigma) \mid {}^i c_j)$$

where $p({}^i s \in \Im({}^i \Sigma) \mid {}^i c_j) = 1 - p({}^i s \in {}^i L \mid {}^i c_j)$.

The net effect of first partitioning in the constraints, then in the lexicalized vs. built distinction is to assign more precise probabilites to each, and to control overgeneration in the building mechanism by partitioning the examples from which the system has to learn the subgrammars.

6. Subgoaling on the sintagmatical problem

Let us define a string generation mechanism to describe the language of valid built signs for constraint ${}^i c_j$, for example a weighted finite state machine [8]:

$$(16)\quad {}^{i-1}A({}^i c_j) = \langle P, {}^{i-1}K, {}^{i-1}\Sigma, {}^{i-1}\kappa_{0(j)}, {}^{i-1}F_j, {}^{i-1}\delta_{0(j)} \rangle$$

where P is the probability weight semiring, ${}^{i-1}K$ the set of building states from an initial state ${}^{i-1}\kappa_{0(j)}$,, ${}^{i-1}\Sigma$ the alphabet to build sequences from, ${}^{i-1}F_j$ a weighted final function for those states of that have built a valid sequence and ${}^{i-1}\delta_{0(j}$ a transition function from states and components to signs, all conditioned on the constraint ${}^i c_j$. Such automaton can easily be inferred from the analyses of the treebank as a Finite Tree Acceptor. Thus:

$$(17)\quad p({}^i x, {}^i s \mid {}^i x_0, {}^i c_j, {}^i s \in \Im({}^i \Sigma)) = p({}^i x, {}^i s \mid {}^i x_0, A({}^i c_j))$$

For each valid hypothesis we have to estimate two probabilities: the probability of constructing a sequence in the automaton or subgrammar, a cumbersome one, and the probability that once obtained the sequence, it represents a valid unit in the upper level:

$$(18)\quad p({}^{i+1} x, {}^{i+1} s \mid {}^{i+1} x_0, {}^i A({}^{i+1} c_j)) =$$
$$p({}^i \chi_0, {}^i \kappa_0 \mid {}^{i+1} x_0, {}^i A({}^i c_j)) \cdot$$
$$\cdot p({}^i \chi^n{}_1, {}^i \kappa^n{}_1, {}^i \varsigma^n{}_1 \mid {}^i \chi_0, {}^i \kappa_0, {}^{i+1} x_0, {}^i A({}^i c_j)) \cdot$$
$$\cdot p({}^{i+1} x, {}^{i+1} s \mid {}^i \chi^n{}_1, {}^i \kappa^n{}_0, {}^i \varsigma^n{}_0, {}^{i+1} x_0, {}^i A({}^i c_j))$$

where $:p({}^{i+1} x, {}^{i+1} s \mid {}^i \chi^n{}_0, {}^i \kappa^n{}_0, {}^i \varsigma^n{}_1 {}^{i+1} x_0, {}^i A({}^i c_j))$ is the probability that the built unit $({}^i \chi^n{}_1, {}^i \kappa^n{}_1, {}^i \varsigma^n{}_1)$ is really valid and give birth to the signs we are interested in:

$$(19)\quad p({}^{i+1} x, {}^{i+1} s \mid {}^i \chi_n, {}^i \kappa_n, {}^i \varsigma^n{}_1, {}^{i+1} x_0, {}^i A({}^i c_j)) \approx$$
$$p({}^{i+1} x \mid {}^i \chi_n, {}^{i+1} s, {}^{i+1} x_0) \cdot {}^i F_j({}^i \kappa_n, {}^i \varsigma^n{}_1 \mid {}^i A({}^i c_j))$$

$p({}^i \chi_0, {}^i \kappa_0 \mid {}^{i+1} x_0, {}^i A({}^i c_j))$ is the final case for the recursion: the probability of starting exploring the automaton. As there is only one starting state for each automaton, and a single observation stream, this probability equals one in our model.

$p({}^i \chi^n{}_1, {}^i \kappa^n{}_1, {}^i \varsigma^n{}_1 \mid {}^i \chi_0, {}^i \kappa_0, {}^{i+1} x_0, {}^i A({}^i c_j))$ is the probability of reaching a particular construction and exploration state with a sequence of units from these initial states. We call calculating these states the *syntagmatical problem.*

7. Solving the syntagmatical problem

We formulate a recursion to solve the syntagmatical problem ending in a particular case as:

$$(20)\quad p({}^i \chi^n{}_1, {}^i \kappa^n{}_1, {}^i \varsigma^n{}_1 \mid {}^i \chi_0, {}^i \kappa_0, {}^{i+1} x_0, {}^i A({}^i c_j)) =$$
$$p({}^i \chi_n, {}^i \kappa_n, {}^i \varsigma_n \mid {}^i \chi^{n-1}{}_1, {}^i \kappa^{n-1}{}_1, {}^i \varsigma^{n-1}{}_1 {}^{i+1} x_0, {}^i A({}^i c_j)) \cdot$$
$$\cdot p({}^i \chi^{n-1}{}_1, {}^i \kappa^{n-1}{}_1, {}^i \varsigma^{n-1}{}_1 \mid {}^i \chi_0, {}^i \kappa_0, {}^{i+1} x_0, {}^i A({}^i c_j))$$

where $p({}^i \chi^{n-1}{}_1, {}^i \kappa^{n-1}{}_1, {}^i \varsigma^{n-1}{}_1 \mid {}^i \chi_0, {}^i \kappa_0, {}^{i+1} x_0, {}^i A({}^i c_j))$ is the base recursion case and $p({}^i \chi_n, {}^i \kappa_n, \varsigma_n \mid {}^i \chi^{n-1}{}_1, {}^i \kappa^{n-1}{}_1, {}^i \varsigma^{n-1}{}_1 {}^{i+1} x_0, {}^i A({}^i c_j))$ is the probability of reaching states $({}^i \chi_n, {}^i \kappa_n)$ with ${}^i \varsigma_n$ after reaching state sequences $({}^i \chi^{n-1}{}_1, {}^i \kappa^{n-1}{}_1)$. As in most modelling we assume a Markov behaviour of these transitions to obtain:

$$(21)\quad p({}^i \chi_n, {}^i \kappa_n, {}^i \varsigma_n \mid {}^i \chi^{n-1}{}_1, {}^i \kappa^{n-1}{}_1, {}^i \varsigma^{n-1}{}_1, {}^{i+1} x_0, {}^i A({}^i c_j)), \approx$$
$$p({}^i \kappa_n \mid {}^i \kappa_{n-1}, {}^i \varsigma_n, {}^i A({}^i c_j))$$

$$p({}^i\chi_n, {}^i\varsigma_n | {}^i\chi_{n-1}, {}^i\kappa_{n-1}, {}^{i+1}x_0, {}^iA({}^ic_j))$$

where $p({}^i\kappa_n, {}^i\varsigma_n | {}^i\kappa_{n-1}, {}^iA({}^ic_j))$ is the transition function in ${}^iA({}^ic_j)$.

8. Subgoaling on the paradigmatical problem

In order to reach states ${}^i\chi_n, {}^i\varsigma_n$ a paradigmatical in this level unit must be obtained and we resort to equation (10) to do so:

$$(22)\quad p({}^i\chi_n, {}^i\varsigma_n | {}^i\chi_{n-1}, {}^i\kappa_{n-1}, {}^{i+1}x_0, {}^iA({}^ic_j)) \approx$$
$$p({}^ix_{0(n-1)}, {}^ic_{0(n-1)} | {}^i\chi_{n-1}, {}^i\kappa_{n-1}, {}^{i+1}x_0, {}^iA({}^ic_j)) \cdot$$
$$p({}^ix_{n-1}, {}^is_{n-1} | {}^ix_{0(n)}, {}^ic_{0(n)}) \cdot$$
$$p({}^i\chi_n, {}^i\varsigma_n | {}^ix_{n-1}, {}^is_{n-1})$$

where $p({}^ix_{0(n-1)}, {}^ic_{0(n-1)} | {}^i\chi_{n-1}, {}^i\kappa_{n-1}, {}^{i+1}x_0, {}^iA({}^ic_j))$ is the probabilitiy that we can find a sign at a lower level of complexity whose syntagmatical function is that expected by the automaton.

$p({}^ix_{n-1}, {}^is_{n-1} | {}^ix_{0(n)}, {}^ic_{0(n)})$ is the base case for the start of the process at equation (10), except that the problem is focused on the sign to transit from ${}^i\chi_{n-1}, {}^i\kappa_{n-1}$ to ${}^i\chi_n, {}^i\varsigma_n$ at a lower level of structural complexity than the initial one.

Finally, $p({}^i\chi_n, {}^i\varsigma_n | {}^ix_{n-1}, {}^is_{n-1})$ is the probability that whatever unit we obtain is suitable for carrying out the transition. We will suppose that ${}^ix_{n-1}, {}^is_{n-1}$ carries enough information to make this probability equal to 1.

The recursion is mainly sketched in the meaning domain through the four meaning levels we are using, - sentence, clause, phrase and word -, but must also proceed in the acoustic level – word sequences, words and phonemes -. This model is exactly the same but needs a more cumbersome formulation for separate delving in each domain.

The whole process terminates in equation (10) because lexemes and phonemes are primitives for our system and on reaching those levels no more recursions are spawned.

4. LANGUAGE MODELLINS AS PARSING

In view of all this, the construction of the language model must entail the parsing of the task corpus to build a tree bank of analyses. In this process, we can use the same framework for speech recognition as we would for parsing: strings of acoustic observations (letters) are the phonetic (orthographic) component evidence from which a full hypothesis (analysis) should be obtained, the difference being the quality of the observations – perfect and partially segmented for the parsing case and error-bearing and unstructured in the recognition case – and the comparison functions for phonetic observations and prototypes proposed by the building process - more flexible for the recognition case. Thus, recognition is conceived as parsing inaccurate input.

However, in our integrated model, recognition and language modelling run parallel because we have accepted that meanings and significants are not independent domains but the different faces of the same coin, - as the saying goes -, so language modelling should in fact be a complex task parsing-model reestimation process over a model such as our own.

This renders traditional language modelling evaluation parameters, like perplexity [1], inadequate, or at least incomplete. Our choice of measure for the time being is sentence likelihood as obtained after analysis completion, but this does not take into consideration coverage or complexity of the task questions as was our initial intention.

5. FURTHER WORK

In pursuing language modelling as language parsing we aim at a better integration with searching algorithms: after the model is evaluated it will be searched with a suitable search algorithm like that of [7]. After the integration we plan to move onto automated learning of the model.

6. REFERENCES

1. Jelinek, F., Mercer, R.L., and Roukos, S. “Principles of Lexical Language Modeling for Speech Recognition”, in Furui, S., and Sondhi, M.M., eds. Advances in Speech Signal Processing. Marcel Dekker, pp. 651-699. New York, Basel: 1991.

2. Bahl, L.R. et al. “A Tree-Based Statistical Language Model for Natural Language Speech Recognition” in *IEEE Trans. ASSP* 37; 1001-1008, 1989.

3. Brown, F.P., et al. “Class-Based n-gram Models of Natural Language”, *Computational Linguistics*, 18(4), pp. 467-479, 1992.

4. Heeman, P.A. *Speech Repairs, Intonational Boundaries and Discourse Markers: Modelling Speakers’ Utterances in Spoken Dialog*. Ph.D.Th. Univ of Rochester, USA, 1997.

5. Brugnara, F., and Federico, M. “Dynamic Language Models for Interactive Applications”, in *Procs. EUROSPEECH’97*, pp. 2751-2754, ESCA, Rhodes: 1997.

6. Ferreti, M., Maltese G. And Scarci, S. “Language Model and Acoustic Model Information in Probabilistic Speech Recognition” in *Procs of the ICASSP*, IEEE: 1989, pp. 707-710.

7. Valverde, F.J. and Pardo J.M. “ A Multi-level Lexical-semantics Based Language Model Design for Guided, Integrated Continuous Speech Recognition”, in *Procs. of the ICSLP’96*. Vol 1, pp. 224-227. Philadelphia, 1996.

8. Pereira, F.C.N, and Riley, M.D., “Speech Recognition by Composition of Weighted Finite Automata”, in Roche, E., and Schabes, Y., *Finite State Language Processing*, Ch. 15, Mit Press, pp. 431-453, Cambridge, 1997.