Generation of Extended Bilingual Statistical Reports

L. Iordanskaja, M. Kim, R. Kittredge, B. Lavoie and A. Polguère
CoGenTex Inc.
810 rue Champagneur, suite 210
Montreal H2V 4S3, Quebec, Canada

1 Introduction

During the past few years we have been concerned with developing models for the automatic planning and realization of report texts within technical sublanguages of English and French. Since 1987 we have been implementing Meaning-Text language models (MTMs) [6, 7] for the task of realizing sentences from semantic specifications that are output by a text planner. A relatively complete MTM implementation for English was tested in the domain of operating system audit summaries in the Gossip project of 1987-89 [3]. At COLING-90 a report was given on the fully operational FoG system for generating marine forecasts in both English and French at weather centres in Eastern Canada [1]. The work reported on here concerns the experimental generation of extended bilingual summaries of Canadian statistical data. Our first focus has been on labour force surveys (LFS), where an extensive corpus of published reports in each language is available for empirical study. The current LFS system has built on the experience of the two preceding systems, but goes beyond either of them [1]. In contrast to FoG, but similar to Gossip, LFS uses a semantic net representation of sentences as input to the realization process. Like Gossip, LFS also makes use of theme/rheme constraints to help optimize lexical and syntactic choices during sentence realization. But in contrast to Gossip, which produced only English texts, LFS is bilingual, making use of the conceptual level of representation produced by the planner as an interlingua from which to derive the linguistic semantic representations for texts in the two languages independently. Hence the LFS interlingua is much "deeper" than FoG's deep-syntactic interlingua. This allows us to introduce certain semantic differences between English and French sentences that we observe in natural "translation twin" texts. LFS is based on a much more detailed text planning process than was attempted earlier, and results in texts of much greater length and complexity. For example, sentence order within certain parts of statistical texts depends on data salience, therefore requiring locally dynamic text planning. Text planning also includes tests that allow for appropriate use of certain quantifier expressions (e.g., *all, most*), evaluative words such as *also* and *only*, and intrasentential pronominalization.

LFS also incorporates some substantial extensions in our use of the Meaning-Text framework. First, it makes more use of lexical functions (cf.[8]), the mechanism in MTMs that allows computation of appropriate collocations and semantically related lexemes needed in paraphrasing and in conflict resolution during generation. Second, the grammar is more extensive, covering important types of conjunction and ellipsis.

Generation in the domain of employment statistics is not new. Roesner's Semtex system [10] produced German (and later, English) summaries of such data that are remarkably similar in style as well as content to our own. The difference lies in our use of a powerful linguistic model that promises to simplify the problem of scaling up the generator to more complex and varied texts, or extend them to other varieties of text. Furthermore, the LFS project is using feedback from domain experts to refine the rules used in both text planning and realization.

2 Text Planning for Statistical Reports

Our approach to planning statistical reports is similar to that used on the Gossip project [2]. A "conceptual frame" tree schema is instantiated with input data to provide an initial characterization of the intended content of the reports. Input data for the employment domain is in the form of relational ta-

[1] The LFS system is being developed by CoGenTex Inc. under contract 36902-0-0749/01-XAF with Communications Canada, Canadian Workplace Automation Research Centre. The first four authors have current academic affiliations with the Université de Montréal. Polguère is now at the National University of Singapore.

bles which provide numerical values for employment, unemployment, participation rate, etc., broken down by age, sex, region and industry, for the current reporting period (e.g., month), as well as for previous comparison periods (e.g., preceding month, one year ago, etc.). The instantiated tree gives a preliminary hierarchical structure for the future text, and provides a framework for further processing on the content to determine the details of text structure. For example, comparisons of employment changes in various labour force groups will lead to ordering messages (future clauses) so as to highlight the most significant changes. The tree structure is traversed and modified as a part of this process. The conceptual text tree also carries annotations of theme and rheme specifications which will constrain the set of possible texts which can be derived from it.

An important part of text planning is the identification of messages which can be grouped together into structures which will give rise to single sentences. This includes conjoining two messages with identical theme to give marked structures that will produce linguistic conjunction and subject pronominalization later as in:

(1) *Employment increased by 20,000 among women while it decreased slightly among men.*

Conceptual conjunction includes checking and marking similarities in thematic elements that may later lead to ellipsis, as in (2), with the possible introduction of lexical functions such as in (3)[2].

(2) *Employment increased by 5000 in Manitoba, by 10,000 in Alberta and by 15,000 in Ontario.*

(3) *For the week ended November 18, 1989, the seasonally adjusted level of employment was estimated at 12,518,000, up 32,000 from October.*

It has been noticed [5] that certain types of report texts have complex internal dependencies that put special demands on the planning mechanism used. In particular, top-down expansion of rhetorical operators is inadequate for generating statistical reports in our domain. Our planning approach, by making use of the power of arbitrary tests and operations on tree schemata, allows us to adequately represent the cross-serial dependencies found among the pieces of content of these reports. However, a more general, but appropriately constrained language for report planning seems to be a desirable goal for future research.

3 Interlingual Representation

Published bilingual reports in our domain occasionally exhibit deep differences between corresponding English and French sentences, as in (4a) and (4b):

(4a) *Employment remained virtually unchanged.*

(4b) *L'emploi a peu varié.*
["Employment changed little."]

Not only are the surface syntactic structures incomparable in this case, but they cannot be easily related on the level of linguistic semantics, because their semantic predicates are dissimilar. We have therefore chosen to use a **conceptual** interlingua (the output of the text planning process) in order to derive separate semantic net representations of the sentences in each language. Hence the sentences (4a) and (4b) are derived from non-isomorphic Meaning–Text semantic networks, which allow us to fully represent the two languages' different "viewpoints" on the same conceptual material.

4 Realizer Design

Grammatical realization in the LFS system is the process by which the semantic nets produced by the planner for the incipient sentences are converted into surface sentences of each language. Our realizer for English is based largely on the general Meaning–Text sentence realizer used in Gossip, with some additions to cover structures found in statistical texts. A comparable realizer for French has been built for LFS. As in the case of Gossip, we use four main linguistic levels of representation between conceptual structures and texts: semantic nets (SemR), deep syntactic dependency trees (DSyntR), surface syntactic dependency trees (SSyntR) and morphological strings (MorphR). For each language, the first linguistic operation requires searching the semantic net for a given sentence to determine the communicatively dominant node. This search is constrained by the theme/rheme specifications which the SemR inherits from the conceptual structure (see [9, 3]). The second operation consists of "replacing" single or complex (configurations of) meaning-bearing nodes in the semantic network by actual lexemes of the language, and replacing semantic features on those nodes by grammatical features which will be attached to the nodes of the future deep-syntactic tree. These operations lead to a reduced semantic graph (RSemR), which is intermediate between SemR and DSyntR. In fact, the SemR is not modi-

[2] The lexeme *up* is the value of the lexical function Adv_1 applied to the verb *increase*.

fied, but rather it is used as a blueprint for building the RSemR, just as each subsequent representation is built by mapping rules from its ancestor representation.

The production of the DSyntR tree out of the RSemR, called "arborization", entails the mapping of predicate-argument relations to deep syntactic relations using information about potential dominant nodes of the RSemR and grammatical features.

The SSyntR is built by mapping deep-syntactic nodes and relations into their surface-syntactic counterparts. Single DSyntR nodes corresponding to phrasemes (i.e., locutions) give rise to syntactic subtrees in SSyntR, and some grammatical lexemes are introduced, including auxiliary verbs, articles and syntactically motivated prepositions.

The next mapping, to MorphR structure, determines word order and all syntactically motivated morphological features. A final operation produces actual text by computing the final (graphical) wordforms based on the morphological features attached to lexemes in MorphR.

5 Lexical Functions

The sublanguage of statistical summary reports shows a certain amount of variation in the syntactic structure and lexical choices used to express a given content. We have used lexical functions to implement this paraphrastic variation in a systematic way within our Meaning-Text models, much as was done in Gossip [3]. Briefly stated, lexical functions (LFs) can be considered abstract meanings which have different lexical values depending on their argument lexemes. LFs provide a way of delaying some idiosyncratic lexical realizations until after major syntactic choices have been made. They also allow us to formulate very general paraphrase rules.

Our statistical reports, with their emphasis on numerical changes and comparisons of change, provide an excellent opportunity to use lexical functions, such as **Magn** ("intensifying" word), $\mathbf{S_0}$ (action nominal) and $\mathbf{Oper_1}$ (agent-oriented support verb). For example, sentence (6) can be calculated to be a paraphrase of (5):

(5) *Employment decreased sharply in October.*

(6) *Employment showed a sharp decrease in October.*

A general paraphrase rule states that a verbal lexeme (here, *decrease*), can be paraphrased by a syntactic construction where the new verb (i.e., *show*) is the value of $\mathbf{Oper_1}$ operating on the nominalization (i.e., $\mathbf{S_0}$) of the old verb. This computation, carried out by successively looking up LF values in argument word lexical entries, derives the new verb *show* by functional composition. In the derived paraphrase sentence (6) this new verb takes as its syntactic object the nominalization of the old verb. In a separate operation which if factored out of the paraphrase operation, the lexical value of the intensifier *sharp* is computed via the lexical function **Magn** operating on lexeme *decrease*. It is simpler to delay its evaluation until after the change in grammatical category of the head word. The paraphrase rule which relates the two verbal constructions of (5) and (6) can be stated using only lexical functions, lexical class (part-of-speech) symbols and grammatical relations, without reference to specific lexical items.

In addition to the above "well-known" lexical functions, our domain also makes use of **Syn** (synonym), **AntiMagn** (diminutive modifier), $\mathbf{Loc_{in}}$ (locative preposition), $\mathbf{Adv_1}$ (locative adverb) and several more "exotic" ones. Most LFs used in our system are introduced during the mapping from RSemR to DSyntR. Exceptions include **Syn**, which is used during reduction of SemR to RSemR. When there are two semantic nodes with identical lexemic meanings, the realizer uses **Syn** to lexicalize one differently from the other by finding a synonym.

6 Implementation and Future Directions

The LFS system is implemented in Quintus Prolog on Sun 4 workstations. Adaptations to several specific varieties of employment reports have been carried out, including the multi-paragraph general summary reports for English and French, given below in §6.1 and §6.2 respectively. The approach outlined here is now being extended to produce other varieties of statistical reports, dealing with different kinds of data (e.g., retail trade summaries). The user interface, which currently allows various choices from among a set of options, is being made more flexible and dynamic by tying the choices more directly to the tree schemata that guide the planning process.

Until now, LFS paraphrasing capability has been implemented only for cases where variation is needed to avoid repetition within a given sentence. The next step, now in preparation, is to enforce variation over longer stretches of text such as whole paragraphs.

6.1 Sample English output

COMMENTARY

Overview

Estimates for November 1989 from Statistics Canada's Labour Force Survey show that the seasonally adjusted level of employment rose by 32000 and that the level of unemployment increased by 30000. The unemployment rate increased by 0.2 to 7.6.

Employment

For the week ended November 3, 1989, the seasonally adjusted level of employment was estimated at 12568000, up 32000 from October. The increase was concentrated among women aged 25 and over. The employment / population ratio remained virtually unchanged (62.1).

Employment among women aged 25 and over rose by 44000 and their employment / population ratio increased by 0.5 to 52.3.

Employment among men aged 25 and over fell by 12000 and their employment / population ratio decreased by 0.3 to 72.5.

Part-time employment increased by 25000. The increase was evenly distributed between men and women.

Full-time employment remained virtually unchanged. An increase among women was offset by a decrease among men.

The level of employment fell by 10000 in agriculture, by 12000 in transportation, communication and other utilities and by 12000 in primary industries other than agriculture. The level of employment rose by 68000 in services and by 20000 in trade. The level of employment remained virtually unchanged in the other sectors.

The level of employment rose by 11000 in Quebec, by 8000 in Alberta, by 6000 in British Columbia and by 5000 in Ontario. The level of employment remained virtually unchanged in the other sectors.

Unemployment and Participation Rate

The seasonally adjusted level of unemployment was estimated at 1032000 for November 1989, up 30000 from October. The unemployment rate rose by 0.2 to 7.6 and the participation rate increased by 0.3 to 67.2.

The increase in unemployment was concentrated among men aged 25 and over.

Unemployment among men aged 25 and over increased by 24000 while unemployment remained virtually unchanged among women aged 25 and over.

The unemployment rate among men aged 15 to 24 increased by 0.7 to 12.9.

The participation rate among men aged 15 to 24 increased by 0.5 to 73.4 and the participation rate remained virtually unchanged among women aged 15 to 24.

The seasonally adjusted level of unemployment remained virtually unchanged in most provinces. The level of unemployment increased only in Ontario (+ 24000).

6.2 Corresponding French output

COMMENTAIRE

Aperçu

Les estimations tirées de l'enquête de Statistique Canada sur la population active pour novembre 1989 indiquent que le niveau désaisonnalisé de l'emploi a augmenté de 32000 et que le niveau du chômage a augmenté de 30000. Le taux de chômage a augmenté de 0.2 à 7.6.

Emploi

Pour la semaine se terminant le 3 novembre 1989, le niveau désaisonnalisé d'emploi est estimé à 12568000, en hausse de 32000 par rapport à octobre. La hausse a principalement touché les femmes de 25 ans et plus. Le rapport emploi / population n'a pratiquement pas varié (62.1).

L'emploi chez les femmes de 25 ans et plus a augmenté de 44000 et le rapport emploi / population chez celles de 25 ans et plus a augmenté de 0.5 à 52.3.

L'emploi chez les hommes de 25 ans et plus a diminué de 12000 et le rapport emploi / population chez ceux de 25 ans et plus a baissé de 0.3 à 72.5.

L'emploi a temps partiel a augmenté de 25000. La hausse s'était également répartie entre les hommes et les femmes.

L'emploi à temps plein n'a pratiquement pas varié. Une hausse chez les femmes a été compensée par une baisse chez les hommes.

Le niveau d'emploi a diminué de 10000 dans le secteur de l'agriculture, de 12000 dans celui des transports, communications et autres services publics et de 12000 dans les industries primaires autres que l'agriculture. Le niveau d'emploi a augmenté de 68000 dans les industries de services et de 20000 dans le secteur du commerce. Le niveau d'emploi n'a pratiquement pas varié dans les autres secteurs.

Le niveau d'emploi a augmenté de 11000 au Quebec, de 8000 en Alberta, de 6000 en Colombie-Britannique et de 5000 en Ontario. Le niveau d'emploi n'a pratiquement pas varié dans les autres provinces.

Chômage et taux d'activité

Le niveau désaisonnalisé de chômage est estimé à 1032000 pour novembre 1989, en hausse de 30000 par rapport à octobre. Le taux de chômage a augmenté de 0.2 à 7.6 et le taux d'activité a augmenté de 0.3 à 67.2.

La hausse du chômage a principalement touché les hommes de 25 ans et plus.

Le chômage chez les hommes de 25 ans et plus a augmenté de 24000 alors que le chômage n'a pratiquement pas varié chez les femmes de 25 ans et plus.

Le taux de chômage chez les hommes de 15 à 24 ans a augmenté de 0.7 à 12.9.

Le taux d'activité chez les hommes de 15 à 24 ans a augmenté de 0.5 à 73.4 et le taux d'activité n'a pratiquement pas varié chez les femmes de 15 à 24 ans.

Le niveau désaisonnalisé de chômage n'a pratiquement pas varié dans la plupart des provinces. Le niveau de chômage a augmenté seulement en Ontario (+ 24000).

References

[1] Bourbeau, L., D. Carcagno, E. Goldberg, R. Kittredge and A. Polguère (1990) "Bilingual Generation of Weather Forecasts in an Operations Environment", *Proceedings of the 13th International Conference on Computational Linguistics*, vol.3, pp. 318–320.

[2] Carcagno D. and L. Iordanskaja (1989) "Content Determination and Text Structuring in GOSSIP", *Extended Abstracts of the Second European Workshop on Natural Language Generation*, Edinburgh.

[3] Iordanskaja, L., R. Kittredge and A. Polguère (1991) "Lexical Selection and Paraphrase in a Meaning-Text Generation Model" in *Natural Language Generation in Artificial Intelligence and Computational Linguistics* (C. Paris, W. Swartout and W. Mann, eds.), Kluwer Academic Publishers, pp.293–312.

[4] Kittredge R., L.Iordanskaja and A. Polguère (1988) "Multi-Lingual Text Generation and the Meaning-Text Theory", *Proc. of the 2nd International Conf. on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Carnegie-Mellon University.

[5] Kittredge R., T. Korelsky and O. Rambow (1991) "On the Need for Domain Communication Knowledge", *Computational Intelligence*, 7(4): 305–314.

[6] Mel'čuk I. (1981) "Meaning-Text Models", *Annual Review of Anthropology*, vol.10, pp.27–62.

[7] Mel'čuk I. and N. Pertsov (1987) *Surface Syntax of English*, Benjamins, Amsterdam.

[8] Mel'čuk I. and A. Polguère (1987) "A Formal Lexicon in the Meaning Text Theory (or, how to do lexica with words)", *Computational Linguistics*, 13(3–4): 261–275.

[9] Polguère, A. (1990) *Structuration et mise en jeu procédurale d'un modèle linguistique déclaratif dans un cadre de génération de texte*, Ph.D. thesis, Université de Montréal.

[10] Roesner, D. (1987) "The Automated News Agency: SEMTEX – A Text Generator for German" *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics* (G. Kempen, ed.), Martinus Nijhoff Publishers, pp.133–148.