An Intelligent Terminology Database as a Pre-processor for Statistical Machine Translation

Michael Carl and **Philippe Langlais**
RALI / DIRO / Université de Montréal
Montréal (Québec), Canada
email:{carl;felipe}@iro.umontreal.ca

Abstract

In a recent study Langlais (Langlais, 2002) has shown that the output of a Statistical Machine Translation (SMT) system deteriorates significantly the more the new text differs from the text the system has been trained on. Langlais shows that bilingual terminological databases are resources that can be taken into account to boost the performance of the statistical engine. This paper extends the notion of 'terminological databases' to an Intelligent Terminological Database (ITDB) capable to detect and reduce terms and their variants and to re-generate the authorized target language terms. The paper discusses the aims and the architecture of the ITDB and evaluates its integration with a SMT system.

1 Introduction

SMT mainly became known to the linguistic community as a result of the seminal work of Brown et al. (1993). Since then, many researchers have invested effort into designing better models than the ones proposed in the aforementioned article and several new exciting ways have been proposed to attack the problem[1].

In a recent paper Langlais (2002) investigated how a statistical engine behaves when translating a very domain-specific text far different from the corpus used to train both the translation and language models used by the engine. Langlais measured a significant drop in performances mainly due to `out-of-vocabulary words` and specific terminology that the models handle poorly. He then proposed to overcome the problem by opening the engine to available (non statistical) terminological resources. This contrasts to a previous approach of (Brown et al., 1993) who develop a statistical model of a bilingual dictionary which is then integrated with training text. Both authors find, however, that terminological databases are resources that boost the performance of a statistical translation engine.

With the possibility to introduce prior knowledge resources into SMT it becomes also interesting to explore their linguistic modeling and to investigate the adaptability of SMT systems to different domains. In this paper, we investigate a possibility to integrate an Intelligent Terminological Database (ITDB) as a pre-processor for an SMT system. This ITDB has the main advantage over simple lists of terms which were used in (Langlais, 2002) as to recognized terminological variants.

Terminological variants are cumbersome in every MT System as they introduce ambiguities which have to be resolved during translation. On the one hand side it is unrealistic and undesirable to list every possible variant in a terminological lexicon. On the other hand, the appropriate target language term has to be generated by the MT engine. In order to overcome this gap, the ITDB follows an abductive approach: a number of possible variants are abduced in a pre-processing step from a list of authorized term translations. The variants and the authorized terms are stored in a database which is consulted at run-time of the tool[2]. The idea being that variants in the source text can thus be traced back to their authorized form and translated properly.

The ITDB is an enhanced version of a ter-

[1] See for instance (Och and Ney, 2000) for a comparison of several translation models.

[2] Variants and Terms are stored in an under-specified format such that the size of the database increases much more slowly than the number of abduced variants. For a more detailed discussion see (Carl et al., 2002).

minology tool described in (Carl et al., 2002) which was adapted and modified here for the bilingual application.

The first part of this paper outlines the aims and architecture of the ITDB. The second part discusses a number of experiments. In section 2, we give an idea of the variants we want to tackle in the ITDB and discuss a number of terminological variants found in an aligned text. Section 3 presents the architecture of the ITDB and section 4 underpins its basic assumptions. In section 5 we show how variants are abduced from a bilingual terminology and sections 6 and 7 report on two experiments.

2 Aim of the ITDB

We have examined an English-French sentence aligned bilingual text form a military domain. The text — which we refer to as SNIPER2 — is a manuals on sniper training and deployment that was used in a previous study (Macklovitch, 1995, cf. section 6).

The text consists of 391 English-French aligned sentences. We have focused on the following phenomena of term variation:

2.1 Variation by Omission

A number of omission variants can be distinguished. The examples (1b,2b) show omission variations for French. In (1b) the expansion pour armes is not specified while in (2b) the type of the lunette is under-specified. Following (Jacquemin, 1996, p. 425), these variants can be said to be in a generic/specific relation.

(1a) *general purpose weapons oil*
↔ huile polyvalente pour armes
(1b) *general purpose weapons oil*
↔ huile polyvalente
(2a) *Unertl telescopic sight*
↔ lunette de tir Unertl
(2b) *Unertl telescopic sight*
↔ lunette Unertl

2.2 Variation by Insertion

Variants by insertion are complementary to omission variants. While in the French term (3b) a new head word tireur is introduced the English term is modified by the additional participle *supported*. In (4b), the English term is permuted and function words are inserted.

(3a) *prone position* ↔ position couché
(3b) *prone supported position*
↔ position du tireur couché
(4a) *rifle butt* ↔ crosse du fusil
(4b) *butt of a rifle* ↔ crosse du fusil

2.3 Synonyms

In addition to insertion and omission, terms also appear as synonyms. As Hamond and Nazarenko (Hamon and Nazarenko, 2001) notice, synonyms may appear in the head and/or in the expansion of a compound. As these different variation processes overlap it becomes particularly difficult to identify the intended meaning. Consider, for instance, the two term-cluster (5a-e) and (6a-e). The terms on the left-hand side in (5a) and (5b) show English variants in their head nouns *telescope* and *scope* while the French variants on the right-hand side in (6a) and (6b) have different expansions tir and visée. There is an English omission variant in (6c) which is translated into a full-form French term and a number of French omission variants (5c,d,e 6d,e). Here it becomes particularly ambiguous and confusing to know whether the full-form authorized translation of French lunette is *spotting telescope* or *telescopic sight*.

(5a)	*spotting telescope*	↔	lunette d'observation
(5b)	*spotting scope*	↔	lunette d'observation
(5c)	*spotting telescope*	↔	lunette
(5d)	*telescope*	↔	lunette
(5e)	*scope*	↔	lunette
(6a)	*telescopic sight*	↔	lunette de visée
(6b)	*telescopic sight*	↔	lunette de tir
(6c)	*sight*	↔	lunette de tir
(6d)	*telescopic sight*	↔	lunette
(6e)	*sight*	↔	lunette

Synonyms and omission variants may thus appear simultaneously, multiplying the 'noise' in translations and aligned texts. It is, however, clear, that one would not like to store all these variants in a bilingual terminology.

3 Architecture of the ITDB

In order to recognize variants of terms and their translations, we have adopted and modified a monolingual terminology tool described in (Carl et al., 2002). The monolingual terminology tool

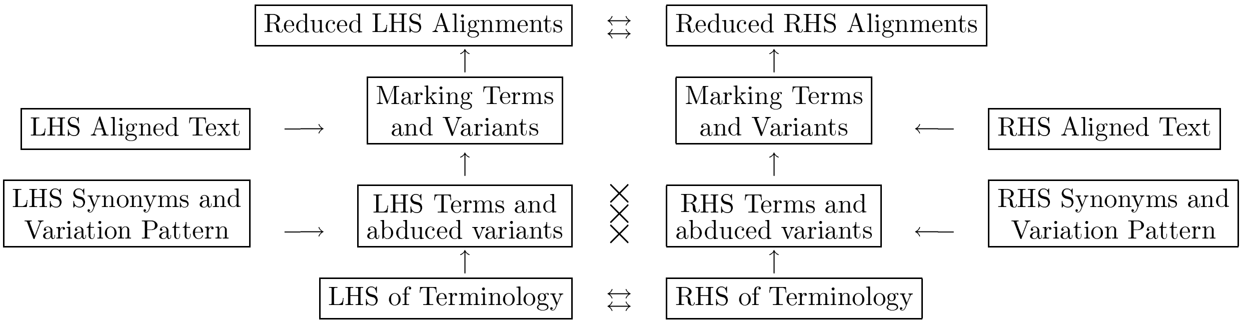
Reduced LHS Alignments
Reduced RHS Alignments
LHS Aligned Text
Marking Terms and Variants
Marking Terms and Variants
RHS Aligned Text
LHS Synonyms and Variation Pattern
LHS Terms and abduced variants
RHS Terms and abduced variants
RHS Synonyms and Variation Pattern
LHS of Terminology
RHS of Terminology

Figure 1: Components of the ITDB

and the ITDB integrate a rule-based formalism KURD and the example-based translation system EDGAR. The modified architecture of the bilingual terminology tool is shown in figure 1. It consists of two symmetrical language sides, a left-hand side (LHS, i.e. English) and a right-hand side (RHS, i.e. French). The architecture in figure 1 is designed to evaluate the performance of the ITDB. A runtime version of the ITDB is shown in figure 3 and discussed in section 7. In this section we describe the different parts of the evaluation architecture.

The ITDB assumes a bilingual terminology (cf. bottom in figure 1). This terminology is either provided by a terminologist or may be automatically acquired. In a study reported in section 6 we use a small bilingual terminology which was manually extracted from the SNIPER2 text. The bilingual terminology contains unambiguous term translations, indicated by the symbol $\leftrightarrow$.

A limited number of variants are generated through the process of "abduction" (discussed below) independently from each language side of a term. Abduction is performed by the rule-based formalism KURD which takes as its input a set of terms, a set of synonyms and a set of variation pattern (see figure 1). The abduced variants are stored in a database together with the original terms and each variant is linked to their authorized forms[3]. Abduction produces $m-to-n$ translations equivalences as indicated by the symbol $\times$.

A bilingual aligned text can now be matched against the database to mark terms and variants in the text as shown in the upper part in figure 1. We use the example-based translation system EDGAR which marks matched sequences of words in the text as variant or term according to the status of the matched database entry. To evaluate coverage and precision of the ITDB (cf. section 6), the marked sequences are reduced to a generic label which are counted and compared in LHS and RHS. In the runtime architecture (cf. figure 3), authorized target language forms of source language terms and their variants are re-generated as discussed in section 7.

4 An Abductive Approach

ITDB recognizes variants of terms by abduction. According to Streiter (Streiter, 2001), abductive reasoning creates hypothesis which are not logically implied by the premises. Unlike deductive reasoning, abduction is not always correct in all reasoning steps. However, abductive reasoning should be "plausible" in a context and yield correct results in the vast majority. Where deductive inference stops in front of gaps, abduction creates new hypothesis which allow to bridge the gap and continue the inference. As an illustration for abductive reasoning, Streiter gives the following example

> "Imagine, you ordered a product and a week later you received a parcel. Using abduction you might assume that this is what you ordered. In order to come to this conclusion you induce from single experiences $\exists(x,y)\ order(x,y) \wedge receive(x,y)$ a hypothesis $\forall(x,y)\ order(x,y) \rightarrow receive(x,y)$ and instantiate x and y with "I" and "product", so that you (safely) assume that you receive your

[3] cf. (Carl et al., 2002) for a more detailed description of this process.

(7) *focusing ring* ↔ bague de mise au point
(8) *eyepiece locking ring* ↔ bague de verrouillage de l'oculaire
(9) *rearward movement* ↔ mouvement arrière

(7_{LHS_0}) *focusing (ring;buckle;collar;light;pneumatic;ripening; ...)*
(7_{RHS_0}) (bague;anneau;aggraver;bouche;boucle; ...) de mise au point
(7_{RHS_1}) bague;anneau;aggraver;bouche;boucle ...

(9_{LHS_0}) *(rearward;recompense;remunerate;pay;compensate) (movement;transport;traffic;trade;...)*
(9_{RHS_0}) (mouvement;émouvoir;transport;trafic;...) (arrière;fond;derrière;sévère;queue; ...)
(9_{RHS_2}) (mouvement;émouvoir;transport;trafic;...) x y (arrière;fond;derrière;sévère;queue; ...)

Figure 2: Abduction of Term Variants

product. "

Essentially, in Streiter's example, from the cooccurrence of two single experiences is inferred an implication containing a universal quantifier in the hypothesis. Abduction consists in reinstantiating the generated hypothesis by appropriate events to draw the desired conclusion.

Mooney (Mooney, 2000) examines the relation between abduction and induction. Although precise definitions of abduction and induction are still somewhat controversial, he finds:

> "In abduction, the hypothesis is a specific set of assumptions that explain the observations of a particular case; while in induction, the hypothesis is a general theory that explains the observations across a number of cases." (Mooney, 2000, p.183)

Mooney applies abductive learning for theory refinement. Theory refinement is the task to make an existing imperfect domain theory consistent with a set of data. For him, abduction is primary useful in generalizing a theory to cover more positive examples. For each individual positive example that is not derivable from the current theory, abduction is applied to determine a set of assumptions that would allow it to be proven.

In a similar way the ITDB detects and reduces terminological inconsistencies. The underlaying assumption in ITDB is that each term in the LHS of an alignment (or in a the source text) has also a translation in the RHS (or in the target text) of that alignment and vice versa. In case a term-translation cannot be detected, the ITDB tries to prove the presence of a variant.

5 Abduction of Term Variants

As outlined in section 3, abduction of term variants in the ITDB presupposes a bilingual terminology. In an evaluation scenario which we shall describe in this section, a bilingual terminology was manually extracted from a text SNIPER2 (cf. section 6). The bilingual terminology contains 168 non-ambiguous term translations where each LHS and each RHS occurs exactly once in the terminology.

Synonyms of the term's content words were generated automatically from a bilingual dictionary by back-and-forth translating[4]. For instance, the terms (7), (8) and (9) in figure 2 were manually extracted from SNIPER2. The terms (7) and (8) contain as their head words the translation *ring* ↔ bague. Back-and-forth translation of French bague yields 29 synonyms while through back-and-forth translation of English *ring* 45 synonyms were generated. In this way, variants (7_{LHS_0}) are recognized as variants of the English term in (7) and variants (7_{RHS_0}) are recognized as variants of the French term in (7). In all, 549 and 650 synonyms were generated from the 168 English and French terms.

In addition to this, a number of variation pattern were used to abduce further variants. Currently, we have the following two simple varia-

[4] We used a general-purpose English-French dictionary with 77016 entries. Lists of English synonyms were obtained by looking up their French translations and then back-translating each of the French translations into English using the same dictionary. In future we consider to use WordNet to obtain synonyms of the term's content words.

tion pattern for French omission (1) and insertion (2) variation :

$$\begin{array}{ll} (1) & N_1 p \ldots \rightarrow N_1 \\ (2) & N_1 Adj_2 \rightarrow N_1 xy Adj_2 \end{array}$$

These variation pattern produced 131 additional variants for the French terms. For instance, the omission variant (7_{RHS_1}) was abduced using variation pattern (1) while insertion variant (9_{RHS_2}) was abduced using variation pattern (2). The tag xy matches any sequence of two words such that, for instance, mouvement vers l'arrière is abduced as a variant of (9). For the English side, only synonyms, but no variation pattern were generated.

While the original terminology contains only $1 - to - 1$ term translations, abduction generates $m - to - n$ translation relations. Thus due to variation pattern (1), French bague is recognized as an omission variant either of term (7) or of term (8). Accordingly, the translations may be English *focusing ring* or *eyepiece locking ring*. Unless another translation is known, the same translation is also abduced for the synonyms: anneau;aggraver;bouche;boucle;.... Abduction thus enables m different French expressions to be translated into n different English terms. However, adding further terms to the terminology will narrow the number of generated translations, as a terminology entry will be preferred over an abduced variant if they describe the same surface string.

6 Coverage and Precision of ITDB

The ITDB was tested on two texts, SNIPER2 and SNIPER3. The texts are an excerpt of an army manual on sniper training and deployment that was used in an other study (Macklovitch, 1995). This corpus is highly specific to the military domain and would certainly prove difficult to any translation engine not specifically tuned to such material.

SNIPER2 and SNIPER3 have 391 and 916 French-English aligned sentences, respectively with an average length of 19 and 22 words in the English LHS and the French RHS. Note that the terminology was also extracted from SNIPER2.

Both language sides of the two texts were passed through the ITDB in two different ways: once only the authorized terms (T) and another time the authorized terms and their abduced variants (T+A) were were marked in both language sides of the aligned texts and retrieved from the output of the evaluation architecture (cf. figure 1). To measure the gain in coverage and precision, we have counted the noise produced in the English and French sides as well as the valid recovered translations equivalences. The table 1 summarizes the results. The row E indicates the noise on the English side of the alignments, i.e the number of matched English expressions which have no correspondences in the French side of the alignment. The row F indicates the noise on the French side. The row $E \leftrightarrow F$ gives the number of valid translations in the alignment.

	SNIPER2		SNIPER3	
	T	T+A	T	T+A
E	131	78	396	343
F	74	220	166	666
$E \leftrightarrow F$	528	638	783	948
PE	0.80	0.90	0.66	0.73
PF	0.87	0.74	0.83	0.59

Table 1: Coverage and Precision of the ITDB
.

In SNIPER2, 659 English terms (i.e. $E \leftrightarrow F + E$) and 602 French terms (i.e. $E \leftrightarrow F + F$) were found using the terminology only, while 716 English and 858 French expressions were matched with abduced variants. Of these matched terms and expressions 528 and 638 were valid translations. This equals a gain of 120% in coverage when using abduced variants.

A similar situation appears for SNIPER3, where 948 valid translations were found using abduced variants compared to 783 translation equivalences for the terminology only. This amounts to an increase of 121% coverage compared to the terminology.

SNIPER	T-T	1-S	S-S	1-T	S-T	2-T
2	528	17	6	80	6	1
3	782	28	10	102	26	0

Table 2: Types and Number of Abduced Translation Equivalences

Analysing the abduced translation equivalences for SNIPER2 and SNIPER3 (cf. table 2), we

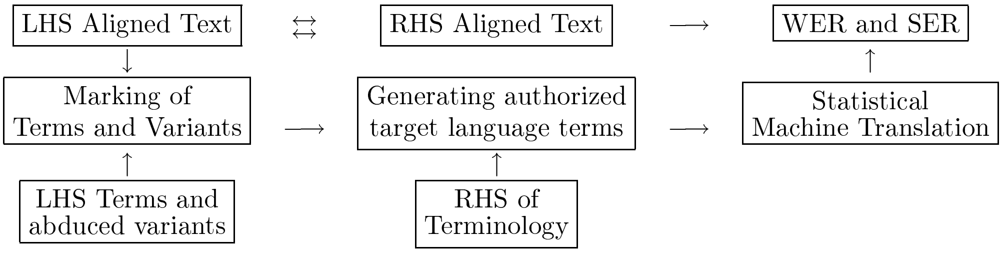
LHS Aligned Text
RHS Aligned Text
WER and SER
Marking of Terms and Variants
Generating authorized target language terms
Statistical Machine Translation
LHS Terms and abduced variants
RHS of Terminology

Figure 3: Runtime architecture of the ITDB.

find the 528 and 783[5] French-English term — term translations (T-T) amongst the the 638 and 948 valid translations. In addition there are 17 and 28 omission-variant — synonym (1-S) translations and 6 and 10 synonym — synonym (S-S) translations and 87 and 127 variant — term (1-T, S-T and 2-T) translations.

By far most of the abduced translation equivalences are due to variation pattern (1) — i.e. the French/English translations 1-S and 1-T — which generates 97 and 130 additional translation equivalences for SNIPER2 and SNIPER3.

However, variation pattern (1) also generates most of the noise in the French alignments. The rows PE and PF in table 1 indicate precision for the English side of alignments calculated as $PE = \frac{E\leftrightarrow F}{E\leftrightarrow F+E}$ and for the French side of alignments calculated as $PF = \frac{E\leftrightarrow F}{E\leftrightarrow F+E}$.

While the precision of the matched English terms increases when matching abduced variants, the precision of the French terms decreases.

	SNIPER2		SNIPER3			
	T	1	T	1	S	2
E	78	0	316	0	27	0
F	68	152	157	492	12	5

Table 3: Origin of Noise in Alignments

Examining the origin of noise in alignments (cf. table 3), more than 2/3 of the matched French sequences which have no correspondence on the English side of the alignment are due to variation pattern 1. As discussed previously (cf. figure 2), each occurrence of the word bouche in the French side of an alignment produces noise if no variant of *focusing ring* or *eyepiece locking ring* was found in the English side of that alignment. This clearly indicates that the variation pattern (1) is too simple — specially if combined with a noisy list of synonyms — which calls for further refinement.

In order to reduce this noise and to extend the coverage of the ITDB, future work will be in line with the methodology of iterative refinement as outlined by Meyer (Meyer, 2001).

7 Integrating ITDB and STM

In a second experiment we have linked the ITDB with an SMT system. The way the ITDB interacts with the SMT engine is depicted in the block diagram of Figure 3. The ITDB identifies and marks terms and their possible variants in the source text (LHS Aligned Text) and regenerates their authorized target language form. The marked and with target language segments enriched text is then passed through the SMT system as described in (Langlais, 2002). While the position of the target term in the French target sentence is determined by the SMT system, its form is generated by the ITDB[6]. in the French alignments . The output of the SMT is then compared with an oracle translation i.e. the RHS Aligned Text.

As in the previous experiment, the architecture was tested on SNIPER2 and SNIPER3, this time in three different settings: without any terminological lexicon, with the terminology (T) and with both the terminology and its abduced variants (T+A). The results of the translation sessions are resumed in Table 4. For practical reasons, we only translated the sentences that contained at most 30 words.

[5] The table 2 shows the number 782: I could not figure out where the missing translation equivalent disappeared ...

[6] Texts were translated from English to French so that the noisy French terms, as reported in the previous section, would not be generated.

	WITHOUT		T		T+A	
corpus	SER	WER	SER	WER	SER	WER
SNIPER2	86.8	82.9	82.6	77.1	82.5	76.6
SNIPER3	91.8	82	91.8	79.4	91.8	79.4

Table 4: SMT Results with the ITDB as a Preprocessor

The performance of our engine was evaluated in terms of `word error rate` (WER) and `sentence error rate` (SER) according to a single oracle translation. The former rate is computed by a classical Levenstein distance; the latter one is given by the ratio of translation that were strictly identical to the oracle translation.

First, the WER measured without terminology is fairly high (more than 82% for both corpora), but in the same range as the ones observed by (Langlais, 2002) in cases where the decoder is faced to texts very different from the ones used at training time. The introduction of the terminology into the engine improves significantly the WER (77% on SNIPER2 and 79.4% on SNIPER3).

Finally, the further introduction of the terminological variants does have a slight positive impact on SNIPER2, but none on SNIPER-3. We have currently no convincing explanation for these findings. We must stress that WER computed over a single oracle translation is probably severe: it may happen that an authorized term translation proposed by the ITDB was not the one present in the oracle translation. This degrades WER even though a correct (i.e. more consistent) translation was produced than was contained in the oracle translation.

8 Conclusion

The paper presents an Intelligent Terminological Database (ITDB), a tool to detect terms and their variants in texts and to retrieve their authorized translations from a bilingual terminology. The paper outlines the architecture of the ITDB and reports on two experiments. The first experiment quantifies the coverage and precision for detecting terminological variants and their translations in aligned texts. In the second experiment, the ITDB is used to translate terms and their variants as a pre-processor for a statistical machine translation system. While the first experiment shows encouraging results, the success of the ITDB as a pre-processor for a statistical machine translation seems more doubtful. A thorough revision and modification of the interaction of both systems is probably in order to fully complement the strength of the two systems. Further experimentation and refinement is also required to reduce the noise produced by the ITDB and to augment its coverage.

References

Didier Bourigault, Christian Jacquemin, and Mairie-Claude L'Homme. 2001. *Recent Advances in Computational Terminology.* John Benjamins Publishing Company, Amsterdam/Philadelphia.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Meredith J. Goldsmith, Jan Hajic, Robert L. Mercer, and Surya Mohanty. 1993. But dictionaries are data too. In *Human Language Technology (HLT)*, pages 202–205, Princeton, NJ, march.

Michael Carl, Johann Haller, Christoph Horschmann, Dieter Maas, and Jörg Schütz. 2002. The TETRIS Terminology Tool. *TAL*, Structuration de terminologie(1).

Thierry Hamon and Adeline Nazarenko. 2001. Detection of synonymy links between terms: Experiment and results. In *in (Bourigault et al., 2001)*, pages 185–208.

Christian Jacquemin. 1996. A symbolic and surgical acquisition of terms through variation. In *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pages 425–438.

Philipe Langlais. 2002. Ressources terminologiques et traduction probabiliste: premiers pas positifs vers un système adaptatif. In *TALN*.

Elliott Macklovitch. 1995. Can terminological consistency be validated automatically ? Technical report, CITI/RALI, Montréal, Canada.

Ingrid Meyer. 2001. Extracting knowledge-rich contexts for terminography. In *in (Bourigault et al., 2001)*, pages 279–302.

Raymond J. Mooney. 2000. Integrating abduction and induction in machine learning. In P. Flach and A. Kakas, editors, *Abduction and Induction*, pages 181–191, Kluwer Academic Publishers.

F.J. Och and H. Ney. 2000. A comparison of alignement models for statistical machine translation. In *COLING00*, pages 1086–1090.

Oliver Streiter. 2001. Treebank Development with Deductive and Abductive Explanation-based Learning: Exploratory Experiments. unpublished draft.