Tailored Feature Extraction for Lexical Disambiguation of English Verbs Based on Corpus Pattern Analysis

Martin HOLUB[1] *Vincent KRÍŽ*[1]
Silvie CINKOVÁ[1] *Eckhard BICK*[2]
(1) Charles University in Prague, Faculty of Mathematics and Physics
Malostranské náměstí 25, Prague
(2) University of Southern Denmark, Campusvej 55, Odense M
{Holub|Kriz|Cinkova}@ufal.mff.cuni.cz, Eckhard.Bick@mail.dk

ABSTRACT
We give a report on a detailed study of automatic lexical disambiguation of 30 sample English verbs. We were drawing on a lexicon of English verb patterns based on the Corpus Pattern Analysis (CPA), which is a novel lexicographic method that seeks to cluster verb uses according to the morpho-syntactic, lexical and semantic/pragmatic similarity of their contexts rather than to associate them with abstract semantic definitions. We have trained several statistical classifiers to recognize these patterns, using morpho-syntactic as well as semantic features. In this paper we mainly concentrate on the procedures for feature extraction and feature selection and their evaluation. We show that tailoring the features to the verbs respectively, as they are implicitly contained in the pattern definitions (explicitly described in the lexicon), has the potential to significantly improve the accuracy of supervised statistical classifiers.

TITLE AND ABSTRACT IN CZECH

Rysy šité na míru anglickým slovesům pro automatickou lexikální disambiguaci pomocí Corpus Pattern Analysis

Předkládáme detailní studii automatické lexikální disambiguace na pilotním vzorku třiceti anglických sloves za použití lexikonu vzorů slovesných užití (patterns), který vychází z Corpus Pattern Analysis (CPA). Tato inovátorská lexikografická metoda namísto na abstraktních definicích jednotlivých významů staví na souhře morfosyntaktické, lexikální a sémantické/pragmatické podobnosti slovesných užití. Natrénovali jsme několik statistických klasifikátorů na rozpoznávání těchto vzorů. Klasifikátory využívají jak morfosyntaktických, tak sémantických rysů. V naší studii se soustředíme na procedury pro extrakci rysů, jejich výběr a jejich evaluaci. Ukazujeme, že rysy na míru uzpůsobené jednotlivým slovesům, jež jsou implicitně obsaženy v definici každého vzoru v lexikonu, mají potenciál významně zvýšit přesnost statistických klasifikátorů s učitelem.

KEYWORDS: English verbs, Corpus Pattern Analysis, supervised lexical disambiguation, tailored feature extraction, machine learning.

KEYWORDS IN CZECH: anglická slovesa, Corpus Pattern Analysis, automatická lexikální disambiguace, rysy šité na míru, strojové učení s učitelem.

Proceedings of COLING 2012: Technical Papers, pages 1195–1210,
COLING 2012, Mumbai, December 2012.

1 Introduction

This study focuses on the lexical semantics of English verbs and its automatic analysis based on contextual hints, morpho-syntactic as well as lexical. It is generally known that the manual annotation of verbs for Word Sense Disambiguation (WSD) tasks has to face the issue of inter-annotator confusion. The commonest verbs are often used to represent several events or aspects of an event at once. For instance, *throwing bread crumbs to the birds* comprises a number of different but interlinked events: propelling an object (typically humans with hands), targeting a propelled object, discarding an object, passing an object to someone else, passing it to an animal as food. Fine-grained lexicons would list all or many of these partial events, since there are good examples of contexts, where one of the aspects is outstanding: throwing missiles, throwing something away/in the sink, throw corn to chickens, etc. In contexts like the one mentioned above, where none of the partial events is dominating, the inter-annotator confusion is almost inevitable, since the instance matches several semantic definitions at once. Using a coarse-grained lexicon, on the other hand, would mean that *throwing darts* and *throwing sour milk in the sink* are similar events, with all the implications for inferencing or translations. Facing this issue, we became fascinated by the Corpus Pattern Analysis, a manual method of sorting corpus concordances according to their morpho-syntactic, lexical and semantic/pragmatic similarity, coined by Hanks (1994). Our current work has been inspired by its implementation, the Pattern Dictionary of English Verbs (PDEV) (Hanks and Pustejovsky, 2005). PDEV as a database has been practically developed at Masaryk University in Brno (Horák et al., 2008) and is publicly available at http://deb.fi.muni.cz/pdev/.

PDEV is a semantic concordance built on yet a different principle than FrameNet, WordNet, PropBank, or OntoNotes: the manually extracted patterns of frequent and normal verb uses are, roughly speaking, intuitively similar uses of a verb that express "in a syntactically similar form" a similar event in which similar participants (e.g. humans, artifacts, institutions, other events) are involved. Two patterns can be semantically so tightly related that they could appear together under one sense in a traditional dictionary. The patterns are *not* senses but syntactico-semantically characterized *prototypes*. Concordances that match these prototypes well are called *norms* while concordances that match them with a reservation (metaphorical uses, argument mismatch, etc.) are called *exploitations* (Hanks, forthcoming). The PDEV corpus annotation indicates the norm-exploitation status for each concordance. Compared to other semantic concordances, the granularity of PDEV is high and thus discouraging in terms of expected inter-annotator agreement. However, selecting among patterns does not really mean disambiguating concordance but rather determining to which pattern it is most similar — a task easier for humans than WSD is. This principle seems particularly promising for verbs as words expressing events, which resist the traditional word sense disambiguation the most.

2 Lexicon of Verb Semantic Patterns

Each lexical entry in the PDEV scheme consists of numbered categories (an example is given in Table 1). Each category consists of a *pattern* and an *implicature*. The pattern represents the morphological, syntactic and lexical characteristics of the verb used in a certain context. The meaning is represented by the implicature. The pattern takes the form of a predication. The pattern-defining verb complements are represented by *semantic types* or *lexical sets*. A lexical set is a list of characteristic collocates, whereas semantic types are items in Hanks' ontology.

Verb	No.	Pattern / Implicature
gleam	1	**[[Physical Object | Surface]] gleam [NO OBJ]** [[Surface]] of [[Physical Object]] reflects occasional flashes of light
gleam	2	**[[Light | Light Source]] gleam [NO OBJ]** [[Light Source]] emits an occasional flash of [[Light]]
gleam	3	**{eyes} gleam [NO OBJ] (with [[Emotion]])** {eyes} of [[Human]] shine, expressive of [[Emotion]]
wake	3	**[no object] [Human] wake ({up}) AdvTime({from} {nightmare | dream | sleep | reverie}) ({to} Eventuality)** the mind of [[Human]] returns at a particular [[Time]] to a state of full conscious awareness and alertness after sleep
wake	4	**pv [phrasal verb] [[Human 1] ^ [Sound] ^ [Event]] wake [[Human 2] ^ [Animal]] ({up})** [[Human 1 | Sound | Event]] causes the mind of [[Human 2 | Animal]] to return to a state of full conscious awareness and alertness after sleep
wake	7	**[Anything] wake [Emotion] ({in} Human)** [[Anything]] causes [[Human]] to feel or become aware of [[Emotion]]
wake	9	**waking * ({up})** [Human|Animal]'s returning to a state of full conscious awareness and alertness after sleep

Table 1: Example patterns defined for the verbs *gleam* and *wake*.

2.1 Pilot Sample English Verbs

We have performed our experiments using a newly developed lexical resource called *VPS-30-En*, recently published by Cinková et al. (2012). *VPS-30-En* (Verb Pattern Sample, 30 English verbs, henceforth VPS) is a pilot lexical resource of 30 English lexical verb entries enriched with semantically annotated corpus samples. VPS is publicly available on the web page http://ufal.mff.cuni.cz/spr/pdev30verbs.[1] The data describes regular contextual patterns of use of the selected verbs in the BNC (2007). VPS has arisen as a practical result of previous studies published by Hanks, drawing on his PDEV, see e.g. (Hanks and Pustejovsky, 2005). VPS contains the verbs showed in Table 2.

VPS is a collection of 30 revised PDEV verbs in which the adjustments of the entries and the original concordance samples were driven by inter-annotator agreement (IAA) findings. The collection was designed as a small sample of PDEV that was revised and cleaned up as a gold-standard data set for statistical pattern recognition.

During the annotation, the annotators got a random 50-concordance sample along with the lexicographer-annotated reference sample and the entry. They matched each random concordance to the categories according to the similarity of implicatures, the similarity of the patterns and, not least, according to the overall similarity of the concordance to the concordance clusters associated with the respective categories.

[1]This language resource has been developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013). In the LINDAT-Clarin repository the VPS data is available under the handle https://ufal-point.mff.cuni.cz/xmlui/handle/11858/00-097C-0000-0005-BF95-B.

Verb	Verb characteristics							Human accuracy				MFC
	Orig. Tagset size	Tagset size	Training set size	Weight %	Frequency group	Perplexity	Fleiss kappa	Annotator AV %	Annotator JT %	Annotator EK %	Average %	Baseline %
access	10	4	300	0.29	C	3.1	0.73	86.3	84.3	86.3	85.6	47.0
ally	8	5	250	0.24	C	3.9	0.73	88.2	80.4	90.2	86.3	47.6
arrive	7	6	250	3.83	B	3.0	0.92	94.1	96.1	94.1	94.8	68.0
breathe	18	7	350	0.60	C	5.1	0.84	94.1	94.1	82.4	90.2	37.7
claim	11	6	500	7.85	A	3.0	0.85	94.1	86.3	92.2	90.9	67.8
cool	16	7	300	0.36	C	5.5	0.88	88.2	82.4	90.2	86.9	27.3
crush	14	9	350	0.27	C	6.9	0.65	82.4	62.8	90.2	78.4	28.9
cry	15	5	250	0.75	B	3.5	0.84	94.1	94.1	88.2	92.2	52.4
deny	12	7	300	3.02	B	5.2	0.69	84.3	68.6	90.2	81.0	44.7
enlarge	6	5	300	0.33	C	2.3	0.62	94.1	66.7	92.2	84.3	76.7
enlist	6	5	300	0.22	C	3.5	0.86	93.8	87.5	91.7	91.0	49.0
forge	14	9	350	0.32	C	7.4	0.64	82.4	72.6	76.5	77.1	26.3
furnish	9	5	300	0.25	C	4.3	0.80	94.1	92.2	74.5	86.9	43.7
hail	10	5	300	0.49	C	2.9	0.83	92.2	98.0	90.2	93.5	67.4
halt	4	4	250	0.54	C	1.8	0.70	94.1	88.2	90.2	90.9	83.6
part	13	9	300	0.24	C	6.2	0.86	94.1	90.2	86.3	90.2	43.0
plough	18	9	250	0.22	C	6.9	0.95	96.1	96.1	92.2	94.8	32.4
plug	14	11	300	0.22	C	8.7	0.72	76.6	80.9	83.0	80.1	31.3
pour	22	10	300	0.57	C	7.9	0.74	90.2	80.4	76.5	82.4	24.3
say	16	6	500	59.3	A	1.9	0.90	96.1	96.1	96.1	96.1	85.2
smash	12	6	300	0.33	C	4.0	0.81	92.2	86.3	90.2	89.5	53.4
smell	11	8	300	0.26	C	5.8	0.88	94.1	86.3	94.1	91.5	36.3
steer	24	14	300	0.27	C	11.1	0.73	80.4	82.4	86.3	83.0	20.3
submit	6	5	250	1.42	B	2.6	0.88	98.0	88.2	96.1	94.1	70.8
swell	25	11	300	0.25	C	9.0	0.82	78.4	80.4	88.2	82.4	21.7
tell	18	9	500	13.5	A	3.8	0.93	98.0	94.1	94.1	95.4	65.2
throw	74	26	1000	2.33	B	16.9	0.65	80.4	62.8	78.4	73.9	22.7
trouble	14	10	300	0.24	C	6.2	0.76	96.1	72.6	88.2	85.6	44.3
wake	11	7	300	0.57	C	4.8	0.78	88.2	82.4	88.2	86.3	45.0
yield	12	10	300	0.93	B	7.4	0.76	86.3	78.4	82.4	82.4	29.0

Table 2: Basic characteristics of the 30 sample English verbs under study. For detailed explanation see Section 3.

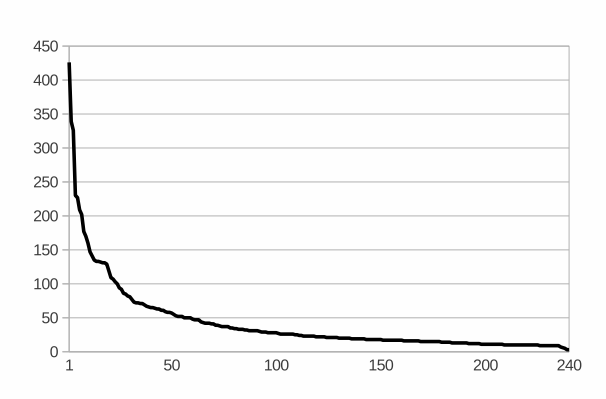

Figure 1: Number of examples of the classified tags in the training data set. The x-axis corresponds to the 240 pattern tags sorted by their frequency in the training data. The value on the y-axis is the number of the training example sentences. For example, only 55 pattern tags have more than 50 examples in our training data.

After each annotation round, IAA was measured and disagreements were manually analyzed. The disagreement analysis was supported by confusion matrices computed for each annotator pair. Provided the annotation of the random sample reached a satisfactory IAA, the disagreements were manually adjudicated by the lexicographer in a spreadsheet table: the lexicographer highlighted evident annotation errors, listed all acceptable values and "one best choice" in a separate column to each concordance. The "one best" annotation was typed back into the user interface as part of the gold standard data set.

The gold standard data set consists of the reference sample and of the adjudication table for the adjudicated sample. As a rule, it consists of 350 concordances (a 250-concordance original sample, one 50-concordance trial sample, and one 50-concordance adjudicated sample).

3 Experimental Data Set

To keep things simple we have neglected the norms and exploitations in this series of experiments. Also, we have, to a significant extent, preserved the classical WSD setup, i.e. both the annotators and the classifier are forced to pick one tag only.

Table 2 shows the most important characteristics of the verbs in the VPS data set. VPS contains about 450 different pattern tags. However, we reduced the number of patterns for classification task and accepted only those with more than 8 occurrences in the training data. The number of pattern tags that we use is 240. The verbs are divided into three frequency groups (A, B, C) according to their frequency in the corpus. Note that 6 most frequent verbs in the data set (*say, arrive, claim, deny, throw* and *tell*) cover 90% of our subcorpus.

The distribution of the number of examples per one tag in the training data set is shown in Figure 1. Unfortunately for most of the 240 tags that should be classified we have a small

Frequency Group	Weight	Perplexity	MFC Average Accuracy %
A	80.7	2.3	73.9 ±0.5
B	12.3	6.5	49.9 ±0.7
C	7.1	5.3	43.7 ±0.5
All	100.0	3.0	68.8 ±0.5

Table 3: Accuracy baselines (the accuracy of a MFC classifier) for all verbs and by frequency groups.

number of training examples (only 91 tags have more than 30 examples).

We also measure the perplexity of the verbs using a standard formula based on the entropy of a probabilistic distribution. The weighted average perplexity of verbs and the baseline accuracy by frequency groups are summarized in Table 3. MFC stands for a most-frequent-case classifier (i.e. the most frequent tag of the training set is used to classify all instances of the test set). The training set was randomly divided into 9 parts to perform a 9-fold cross-validation. In Tables 3 and 6 the AVG accuracy is displayed together with the confidence intervals based on the standard t-test at the significance level $\alpha = 5\,\%$.

The test sets contain 50 multi-annotated instances of each verb. The IAA values (measured as Fleiss' kappa) displayed in Table 2 were calculated using human annotations made independently by 4 annotators. In Table 2 we also indicate the values of "human accuracy", just to illustrate how difficult the classification task is for people.

4 Feature Extraction and Selection

We have identified the feature extraction for machine learning as a central issue, with great impact on the performance of automatic classifiers. Since we used only one-sentence contexts, we dealt only with *local* features. Each data instance to be classified consists of the target verb (TV) and some context words. Therefore the features that describe data instances are based on the observed characteristics of both the TV and the context words. Intuitively it seems to be a good idea to follow the structure of the defined patterns. Therefore we use two kinds of features for machine learning, the morpho-syntactic features and the semantic features. All features used in this study are binary, i.e. have only 0/1 values.

4.1 Types of features and feature sets

4.1.1 Morpho-syntactic features

We have considered three types of morpho-syntactic features: 1) morphological features of the target verb, 2) morphological features of words in a contextual window, 3) syntactic dependencies. By the morpho-syntactic feature set we were seeking to analyze collocates in relevant argument positions, negation and the modifiers of arguments. To alleviate the automatic parser errors, we formulated some features just based on part-of-speech tags. For instance, a (likely) direct object is also encoded as a noun following the target verb in a given window. We were also observing the tense, mood and voice of the target verb. Also, we were taking into account whether the verb is governed by or governs another verb. We were drawing on (Semecký, 2007), adapting the features to English.

First we tuned a model for pattern classification using only morpho-syntactic features, and only then we tried and improved it by using a set of semantic features.

4.1.2 Semantic features

Semantic features capture the semantics of nominal subjects and/or objects of the TV and/or prepositional phrases dependent on the TV, which exactly corresponds with semantic types and/or lexical sets indicated in the pattern definitions. For that purpose we use the system of *semantic prototypes* developed by Bick (first publicly mentioned in (Bick, 1996); the current documentation is available on the Web[2]). In fact, to create semantic features we take only the most coarse-grained level of the semantic prototypes, which is called "umbrella" categories. In Bick's system each of the 150–200 semantic prototypes is assigned to some of 40 umbrella categories, which are grouped into 22 "major umbrellas": *animal, botanical, human, location, vehicle, abstract concepts, actions/events/processes, anatomical, things (concrete/countable), clothes, materials, collectives/parts, domain concepts, features, food, perceptions and feelings, semantic/semiotic, state of affairs, time, tools/machines, units/quantities, weather*.

The semantic prototypes for English are drawing on similar systems for Portuguese and Danish. The original motivation for creating the semantic prototypes, such as <Hprof> (professional human), <tool> or <sem> (semantic/semiotic product) was the polysemy resolution for Portuguese-Danish machine translation in a Constraint Grammar context. Thus, context-driven rules are used to remove or select semantic categories for a given lemma, depending on its syntactic function, inflexion, definiteness and dependency relations. The granularity of the semantic prototypes was chosen to match linguistic usefulness, rather than for its descriptive value as such. Thus, tags should optimally be distinguishable with linguistic tests, such as the combinatorial potential, e.g. which prepositions are typically used with a noun in question, plural vs. mass determiners, or testing verbs: "you can eat it, drink it, write it . . . ". Too low a granularity (high level of abstraction) would reduce the distinctive power, too high a granularity (low level of abstraction) would make it impossible to express general contextual rules and not gain much compared to lexical rules targeting the individual lemma.

4.1.3 Universal and tailored feature sets

We developed and evaluated two kinds of feature sets. While *universal* feature sets are common to all verbs under study, *tailored* feature sets are verb-specific. The procedure for extracting the tailored features is fully automatic and is based on an automatic analysis of the pattern definitions. As we show in this paper, using tailored feature sets enables us to build automatic classifiers with a significantly better performance.

After many experiments we tuned and evaluated 5 models for feature extraction and selection. First, we started with a basic universal model that deals only with morpho-syntactic features (U1). Second, to evaluate the contribution of semantic features, we designed a more advanced universal model that also uses semantic features (U2). Then we focused on using the specific clues contained in the pattern definitions and developed 3 tailored models (T1, T2, and T3) that work with all kinds of features and differ in procedures for feature selection. An overview is given in Table 4.

[2] http://beta.visl.sdu.dk/semantic_prototypes_overview.pdf

Model	Number of features	Relation to verbs	Feature characteristics	Feature selection method
U1	58	universal	morpho-syntactic features only	selection using Decision Trees (Information Gain)
U2	65	universal	morpho-syntactic and semantic features	same as U1
T1	68-106	tailored	the U2 set + features based on pattern definitions	same as U1 and U2 + features tailored to verbs
T2	19-88	tailored	all morpho-syntactic, semantic, and pattern-based features	greedy forward selection to maximize SVM accuracy
T3	29-147	tailored	the union of the feature sets selected by T1 and T2	greedy backward elimination to maximize SVM accuracy

Table 4: Overview of the models used for feature extraction.

4.2 Morpho-syntactic Feature Extraction

Morpho-syntactic features are extracted from sentences using both a morphology analyzer based on the Penn TreeBank (PTB) morphological tagset (Santorini, 1990) and the Stanford parser with its Stanford dependencies representation (de Marneffe et al., 2006). In total we have established 79 fixed binary features and 4 lexicalized features. In fact the lexicalized ones generate a number of binary features, depending on the occurrence of certain auxiliary words (prepositions, particles, and conjunctions) in the training data. All morpho-syntactic features can be divided into 3 groups:

Characteristics of the TV
10 binary features: Passive voice, modality-1 (would, should), modality-2 (can, could, may, must, ought, might), negation, tense (PTB tags: VBN, VBD, VBG, VBP, VB), use in an infinite phrase (outside subject).

Characteristics of the context words that immediately precede or follow the TV
Context is limited to ± 3 words simply by the word order. 9 binary features have been established for each of the 6 closest context words (in total 54 binary features): nominal-like (NN, NNS, NNP, NNPS, DT, PDT, PRP, PRP$, POS, CD), adjective (JJ, JJR, JJS), verbs (VB, VBD, VBG, VBN, VBP, VBZ), modal (MD), adverbial (RB, RBR, RBS, RP, IN), *to* (TO), wh-pronoun (WDT, WP, WP$), wh-adverb (WRB), to_be (lemma = *be*).

Characteristics of the context words that syntactically directly depend on the TV
a) logical subjects (3 binary features: nominal subject, clausal subject, subject in the plural form); b) objects (8 binary features: direct object, indirect object, passive nominal subject, passive clausal subject, clausal complement, complementizer (typically the subordinating conjunction "that" or whether"), any object, any object in the plural form); c) particles (lexicalized); d) adverbials (4 binary features: adverbial modifier, adverbial clause modifier, purpose clause modifier, temporal modifier); e) preposition (lexicalized: prepositional modifier or prepositional clausal modifier); f) markers (lexicalized: subordinating conjunctions different from *that* or *whether*).

4.3 Morpho-syntactic Feature Selection

Feature selection was performed in two steps. First, we filtered out the features with useless value distribution. As a threshold we used the condition that the less frequent (binary) value should be detected in our training data at least 5 times. After that filtering we had 149 binary features. The second step was reducing the remaining feature set in order to "optimize" the classifier performance. After many experiments the following heuristic procedure won. For each of the 30 verbs separately we searched for a small subset of the features with the best performance using a decision tree classifier. We started with only one best feature and then greedily added further best features and tested the classifier performance. When it was not possible to improve the accuracy by adding any of the remaining features, the process stopped. Then all the "best" small sets for all 30 verbs were united and we got an overall feature set containing 58 morpho-syntactic features. We call this model "U1". Its accuracy for different verbs is shown in Table 5 and in Figure 2. We experimentally checked that U1 could be hardly beaten regarding the overall accuracy measured as a weighted average of all 30 verbs.

4.4 Semantic Feature Extraction and Selection

Our universal model U2 deals only with the umbrella prototypes (22 major + 40 subordinated) and observes only the semantic types of subjects and objects. As a starting point, the feature selection procedure takes the overall set of 124 (binary) semantic features (62 for the semantic type of subjects, and other 62 for objects). The final feature selection was done analogously as the previous selection of the morpho-syntactic features. We started with the union of the best 58 morpho-syntactic features and all 124 semantic features. First, we greedily searched for the "best" small feature subset for each of 30 verbs separately. Then we took the union of all small subsets. The result was a set of 65 features consisting of 21 semantic and 44 morpho-syntactic ones. We call this model "U2". Again, its accuracy for different verbs can be seen in Table 5 and in Figure 2. The overall average results for all 30 verbs are given in Table 6.

4.5 Tailored Feature Extraction and Selection

The extraction of the tailored features is based on contextual hints described in the patterns. This process is driven by 1) the presence of a member of a lexical set defined in patterns, 2) the verb forms indicated in patterns, 3) prepositions listed in prepositional phrases described in patterns, 4) particles dependend on the TV, 5) types of object clauses allowed in patterns, 6) the "no_object" attribute defined in patterns.

The tailored models differ in the feature selection method used. The T1 model simply uses the U2 feature set and all tailored featured corresponding to a given verb. The T2 model takes all possible features and greedily selects the best ones to maximize accuracy of an SVM classifier. The most advanced T3 model takes the union of the T1 and T2 feature sets and then reduces the whole set by greedy backward elimination, again to maximize accuracy of an SVM classifier.

5 Supervised Pattern Classification: Model Choice and Tuning

We experimented with several supervised machine learning methods, namely k-Nearest Neighbours (kNN), Decision Trees (DT), AdaBoost.M1 (ADA) based on DT, Support Vector Machines (SVM), and Naive Bayes classifier (NB). Our results are perfectly in line with the observation reported in (Màrquez et al., 2007) that the best results are obtained using SVM or ADA. We also observed that in case of small samples for different patterns, the SVM model tends to be

Verb	Average accuracy						Best tailored model				
	MFC	U1	U2	T1	T2	T3	M	#F	Acc	Imp-B	Imp-U
access	47.0	78.0	77.7	77.4	79.0	79.7	T3	55	79.7	69.5	2.6
ally	47.6	65.5	66.8	67.6	79.6	79.2	T2	54	79.6	67.3	19.2
arrive	68.0	70.8	72.4	76.8	82.0	82.6	T3	41	82.6	21.4	14.1
breathe	37.7	65.7	72.3	76.3	79.4	81.0	T3	41	81.0	114.7	12.0
claim	67.8	82.4	82.8	87.4	80.6	82.6	T1	75	87.4	28.9	5.5
cool	27.3	63.0	63.4	65.4	66.7	67.6	T3	36	67.6	147.1	6.6
crush	28.9	37.1	46.3	50.3	53.4	53.5	T3	56	53.5	85.3	15.5
cry	52.4	72.4	73.6	77.2	78.8	80.4	T3	44	80.4	53.4	9.2
deny	44.7	55.7	60.7	67.7	63.3	63.0	T1	74	67.7	51.6	11.5
enlarge	76.7	82.0	80.7	84.0	82.0	84.8	T3	43	84.8	10.6	5.1
enlist	49.0	74.4	84.4	84.7	89.4	89.9	T3	51	89.9	83.5	6.6
forge	26.3	48.3	52.6	59.7	56.9	58.6	T1	86	59.7	127.2	13.6
furnish	43.7	65.0	69.7	72.0	77.7	79.0	T3	49	79.0	80.8	13.4
hail	67.4	85.0	83.7	85.4	81.7	84.6	T1	73	85.4	26.7	2.0
halt	83.6	85.2	86.8	87.6	88.0	90.9	T3	59	90.9	8.8	4.8
part	43.0	73.0	73.0	72.7	69.0	67.0	T1	74	72.7	68.9	-0.4
plough	32.4	69.3	70.9	73.6	74.0	76.5	T3	44	76.5	135.9	7.9
plug	31.3	51.0	59.4	58.7	58.7	61.7	T3	41	61.7	96.7	3.9
pour	24.3	52.3	55.7	56.7	58.9	63.8	T3	77	63.8	162.1	14.5
say	85.2	90.6	90.6	90.8	86.0	86.9	T1	82	90.8	6.6	0.2
smash	53.4	65.1	69.7	74.3	76.7	77.7	T3	46	77.7	45.7	11.5
smell	36.3	57.0	61.0	63.0	58.3	63.7	T3	37	63.7	75.2	4.3
steer	20.3	40.4	44.0	45.6	49.0	50.6	T3	55	50.6	149.1	15.1
submit	70.8	85.2	85.2	85.6	84.0	86.8	T3	76	86.8	22.6	1.9
swell	21.7	46.6	51.0	57.3	62.0	62.8	T3	45	62.8	189.8	23.2
tell	65.2	75.8	79.2	79.4	79.2	81.2	T3	69	81.2	24.6	2.5
throw	22.7	43.0	42.6	53.7	56.6	56.6	T3	147	56.6	149.3	32.9
trouble	44.3	70.7	69.7	72.4	66.0	65.5	T1	75	72.4	63.2	3.8
wake	45.0	76.7	77.3	77.7	69.7	69.8	T1	75	77.7	72.6	0.5
yield	29.0	46.6	51.9	52.6	55.3	56.0	T3	46	56.0	93.1	7.8

Table 5: Comparison of two universal and three tailored models. The best model is always one of T1, T2, or T3. #F is the number of the features selected by the best model. Imp-B and Imp-U stand for improvement over the baseline and over the best universal model U2, respectively.

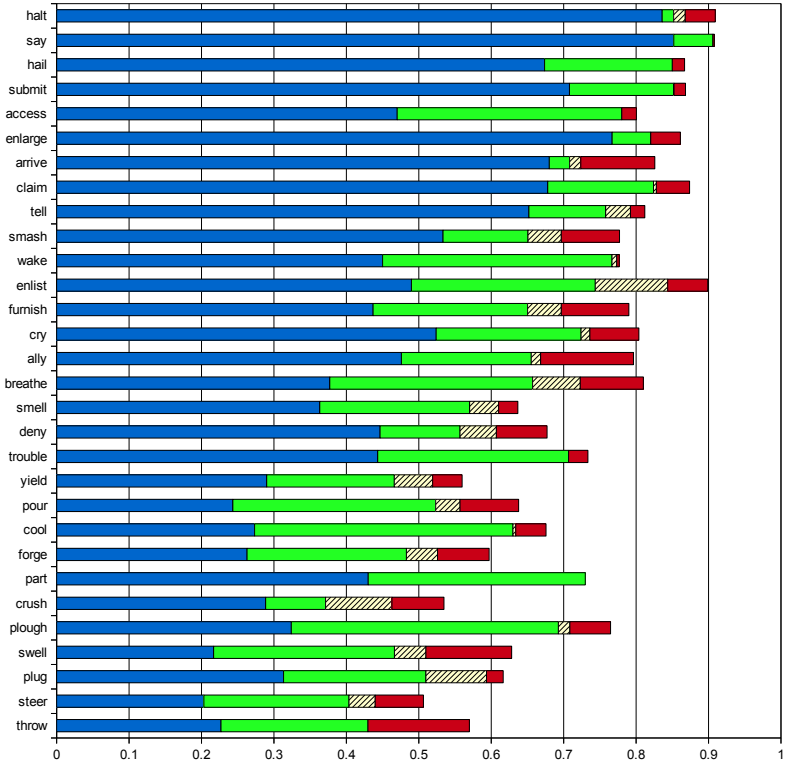

Figure 2: Accuracy improvement over the MFC baseline (in blue): the best models using universal feature sets U1 (green) and U2 (cream), and the best models using tailored feature sets (red).

Frequency Group	Best universal				Best tailored			
	AVG Accuracy %	Improvement %	ERR decrease %	Test Acc %	AVG Accuracy %	Improvement %	ERR decrease %	Test Acc %
A	87.9 ±1.9	10.4	38.2	86.4	88.9	11.8	41.5	83.1
B	63.9 ±2.1	38.3	27.1	58.8	72.3	60.6	45.6	66.6
C	68.7 ±0.9	72.0	42.8	66.9	74.1	87.8	53.4	69.6
All	83.6 ±1.6	18.2	37.1	81.6	85.8	23.1	42.8	80.1

Table 6: Accuracy of the best SVM models. *AVG Accuracy* is the result of the cross-validation test on the training data set. *Improvement* is the percentage difference between the AVG Accuracy and the Baseline Accuracy. *ERR decrease* stands for the percentage difference between the baseline error rate and the error rate of the respective model. *Test Acc* is the accuracy measured on the test sets.

better than ADA (cf. the overview article by Schapire (2003)). Finally we trained our best models (U1, U2, T1, T2, T3) using the SVM method and a grid search approach to parameter optimization. The results described in Table 5 and in Figure 2 show that tailored models almost always clearly outperform the universal ones. However, among the tailored models there is no absolute winner. Unfortunately, we have not yet developed one "best" method for feature selection that would universally lead to the best performance for any verb. Therefore we chose the best tailored model for each verb separately, according to the results of the cross-validation experiments performed on the training data.

6 Evaluation and error analysis

6.1 Universal models

Our overall result is that the "semantically enriched" model U2 slightly outperforms the "only morpho-syntactic" model U1, which can be observed in Table 5. However, only the difference for the low-frequent verb group is statistically significant. In Figure 2 the verbs are sorted according to the increasing perplexity; this figure also shows the decreasing accuracy tendency, which has naturally been expected.

The classifiers did not know the pattern definitions for the respective verbs. We made this decision in order to see to what extent the default feature set would do. This approach, along with the sparsity of our data, resulted in a few systematic errors. The most striking one is the misclassification of patterns that prescribe a participial form of the verb. When these patterns were not frequently assigned, the classifier did not learn the most important feature — that the verb should be a participle that is not used in an obvious passive voice, but is already a transition to an adjective or a noun (e.g. *cooling* in *cooling towers*). The classifier often assigned this tag to concordances in which the verbs did not have the form of a participle.

Another drawback of the classifiers is that they have not been fed a list of phrasemes. As the data is so sparse, idioms either remain unrecognized, or they are interpreted literally. For

Feature type	Frequency	%
All tailored	**283**	**16.1%**
Tailored – lexical sets	94	5.3%
Tailored – prepositions	68	3.9%
Tailored – particles	24	1.4%
Tailored – clauses	15	0.9%
Tailored – verb form	59	3.4%
Tailored – no object	23	1.3%
All semantic	**462**	**26.2%**
Semantic – Subj	118	6.7%
Semantic – Obj	304	17.3%
Semantic – PP	40	2.3%
All morpho-syntactic	**1016**	**57.7%**
All	**1761**	**100.0%**

Table 7: Overview of the structure of features used in best tailored feature sets.

instance the concordance *This organisation happily <ploughs> a furrow totally at odds with the notion of free trade* is interpreted as an agricultural context. This problem goes beyond just idioms, since many patterns with limited collocability are defined by lexical sets — lists of typical collocates. The nouns in these lists are often quite heterogeneous, encompassing several semantic types and the association with a semantic type is irrelevant. For instance *claim credit for something*. The data is too small for such lists to be learned directly.

Another interesting issue is semantic modulation in nouns. For instance, the verb *halt* distinguishes between abstract processes, such as financial crises, and vehicles or human groups in military contexts to be halted. In the following concordance, *advance* is a process, but what is really meant are the men and vehicles advancing. Semantic modulation is a typical cause of annotator confusion, often mimicked by the classifier: *And even Crown Prince Rupprecht, far removed from Verdun, had warned him days before the offensive began that the advance would be <halted> by flanking fire from the Left Bank.*

6.2 Tailored models

To some extent, tailored feature sets are able to provide a remedy for the errors described above. A summary is given in Table 6 to compare our best universal model U2 with the best tailored models (specific for each verb). Although tailored features cause the accuracy increase, Table 7 indicates the fact that the morpho-syntactic features dominate even in the best tailored models.

A glance at the results provided by the tailored models reveals a couple of observations. We have compared the results for the test data with the human confusion matrices and with the overall outcome of the universal models as we have described it in the previous section. We have identified four interesting points:

1) Participial patterns. The universal models did not learn that the confusion between a participial and a regular pattern is only acceptable when the target verb is in a participial form. The tailored models learned this successfully for most verbs where participial patterns occurred. There is only one major exception: cool 11 (participial) – confused for cool 1 (intransitive).

2) Patterns with a different number of objects. In several verbs, e.g. "deny something" (deny 9) confused for "deny somebody something" (deny 10), neither model learned to discriminate according to the number of non-prepositional objects, although the presence of the indirect object was among the features. We suspect this confusion to occur due to parsing errors.

3) Syntactically similar patterns with different implicatures or semantic types. The classifiers are in trouble whenever two pattern definitions are syntactically similar and the only difference lies in the semantic types of the collocates. Although the universal features contain the semantic types, this semantic information is not sufficiently granular. Unlike lexical sets, we do not have any detailed information on which words correspond to which semantic types.

4) Heterogenity of 'u' and 'x' tags. The most systematic error is in many verbs a pattern number assigned to a concordance classified as 'u' or 'x'. We speculate that learning these negative instances is extremely difficult. Their examples in the data are very heterogeneous, and each of them can be more similar to a positive instance, respectively, than they are among one another.

6.3 Future work

As a next step, we would like to exploit the potential of the semantic types determined in the patterns. We need to develop a robust method to "populate" the semantic types with lexical units. Also, we need to gain a better insight into the performance of parsers, since the most important features are inarguably the syntactic ones. A weak spot of the tailored features is that, in *some* cases, our best tailored models slightly overfit the training data (as can be observed in Table 6). So we need to make the proces of feature selection more robust. The main issue, however, that sets the limits on the performance of supervised classifiers seems to be the lack of sufficient amount of reliable training examples.

Conclusion

The two main goals of our research were to evaluate the usefulness of semantic prototypes if we use them directly as features for statistical learning, and to evaluate the power of features tailored to individual verbs and based on automatic analysis of pattern definitions. Our result is in line with previously published studies that usually agree on the fact that the morpho-syntactic features are the most important for statistically-driven semantic disambiguation. Nevertheless, for *some* verbs the use of semantic features plays an important role. The positive impact of tailored features is obvious.

Acknowledgments

This research work has been supported by the Czech Science Foundation (grant project no. P103/12/G084) and partly by the project META-NET (FP7-ICT-2009-4-249119 of the EU and 7E11040 of the Ministry of Education, Youth and Sports of the Czech Republic).

We thank our friends from Masaryk University in Brno for providing the annotation infrastructure and for their permanent technical support. We thank Patrick Hanks for his CPA method, for the original PDEV development, and for numerous discussions about the semantics of English verbs.

References

Bick, E. (1996). Automatic parsing of Portuguese. In *Garcia, Laura Sánchez (ed.), /Anais / II Encontro para o Processamento Computacional de Português Escrito e Falado/*. Curitiba: CEFET-PR.

Cinková, S., Holub, M., Rambousek, A., and Smejkalová, L. (2012). A database of semantic clusters of verb usages. In *Proceedings of the LREC 2012 International Conference on Language Resources and Evaluation*. Istanbul, Turkey.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Hanks, P. (1994). Linguistic norms and pragmatic exploitations, or why lexicographers need prototype theory and vice versa. In F. Kiefer, G. K. and Pajzs, J., editors, *Papers in Computational Lexicography: Complex '94*. Research Institute for Linguistics, Hungarian Academy of Sciences.

Hanks, P. and Pustejovsky, J. (2005). A pattern dictionary for natural language processing. *Revue Francaise de linguistique appliquée*, 10(2).

Horák, A., Rambousek, A., and Vossen, P. (2008). A distributed database system for developing ontological and lexical resources in harmony. In *9th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15. Berlin: Springer.

Màrquez, L., Escudero, G., Martínez, D., and Rigau, G. (2007). Supervised Corpus-Based Methods for WSD. In Agirre, E. and Edmonds, P., editors, *Word Sense Disambiguation*. Springer, 2007.

Santorini, B. (1990). Part-of-Speech tagging guidelines for the penn treebank project. Technical Report MS-CIS-90-47, University of Pennsylvania.

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In Denison, Hansen, Holmes, Mallick, and Yu, editors, *Nonlinear Estimation and Classification*. Springer, 2003.

Semecký, J. (2007). Verb valency frames disambiguation. PhD. Thesis. Technical report, Institute of Formal and Applied Linguistics, Charles University in Prague.