Coarse Semantic Classification of Rare Nouns Using Cross-Lingual Data and Recurrent Neural Networks

Oliver Hellwig
Düsseldorf University, SFB 991
ohellwig@
phil-fak.uni-duesseldorf.de

Abstract

The paper presents a method for WordNet supersense tagging of Sanskrit, an ancient Indian language with a corpus grown over four millenia. The proposed method merges lexical information from Sanskrit texts with lexicographic definitions from Sanskrit-English dictionaries, and compares the performance of two machine learning methods for this task. Evaluation concentrates on Vedic, the oldest layer of Sanskrit. This level of Sanskrit contains numerous rare words that are no longer used in the later language and whose word senses can, therefore, not be induced from their occurrences in other texts. The paper studies how to efficiently transfer knowledge from later forms of Sanskrit and from modern Western dictionaries for this special task of supersense disambiguation.

1 Introduction

The paper discusses experiments in coarse-grained word semantic disambiguation (WSD) for Classical (CS) and Vedic Sanskrit (VS).[1] These experiments are part of a project that deals with the verb-argument labeling of Vedic texts. The project is based on a manual annotation of all 27,104 verbal forms and their main arguments found in the Ṛgveda (ṚV), the core text of the Vedic corpus (Hettrich, 2007).[2] Apart from relating arguments to their governing verbs, the annotation disambiguates case semantic functions such as time or location for the locative, and it assigns a basic word semantic class to each argument (refer to the sample annotation in Fig. 1). The word semantic annotations differentiate between eight classes that include, among others, abstract concepts, humans, and animals. We are planning to use the annotation of the ṚV as training corpus for building a verb-argument labeler that can be applied to other texts of the Vedic corpus.

Several publications on argument and role labeling use word semantic classes or distributional representations of words for modeling selectional preferences of verbs (Wilks, 1975; Che et al., 2010; Yu et al., 2010; Roth and Lapata, 2015). Following this work, we are going to employ WordNet supersenses (WNSS; Ciaramita and Johnson, 2003) of Vedic words as an additional prior in our argument labeling pipeline, both for detecting arguments in unlabeled texts (see the semantic coherence criterion in Laparra and Rigau (2013)), and for assigning appropriate word semantic classes to arguments.

The paper interprets WSD as a sentence classification task, where definitions from bilingual Sanskrit-English dictionaries and sentential contexts serve for predicting word semantic classes of Sanskrit nouns. The paper concentrates on rare nouns, because the vocabulary of Vedic texts contains numerous lemmata that have disappeared in later Classical Sanskrit, so their distributional properties cannot be estimated

[1] Sanskrit can be divided into two historical layers, whose relationship resembles that of Homeric and Classical Ancient Greek, or even the later *koine*. Vedic Sanskrit is one of the oldest Indo-European languages. Its earliest parts may have been composed in the second millenium BCE (Witzel, 1995). Around 350 BCE, the grammarian Pāṇini compiled the grammar Aṣṭādhyāyī, a linguistic overview of a late form of VS, which became the prescriptive standard for CS (Scharfe, 1977). Although the vast majority of Sanskrit texts is written in CS, VS also has produced a sizeable corpus of several million words.

[2] The annotation was performed by a H. Hettrich, and parts of it were later inspected randomly by linguists; personal communication by H. Hettrich.

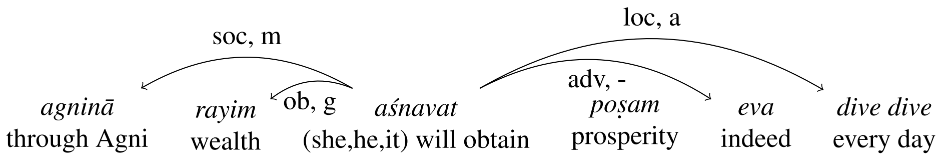

Figure 1: Verb-argument annotation of Ṛgveda 1.1.3 ("He will obtain wealth [and] prosperity through Agni every day."). Labels on the arcs indicate the syntactic functions (soc[iative], ob[ject], loc[ation]) and coarse word semantic classes (m = human, g = object, a = generic expression) of the arguments. If more than one word fits into an argument class, only the first one is annotated.

reliably from the later corpus. We will compare the efficiency of different machine learning models for this task. In addition, the paper pays special attention to the philological setting of WSD. While most NLP studies work with huge, contemporaneous corpora from closed domains (newspapers, Twitter), and can rely on richly annotated data sets, WSD for Vedic and Classical Sanskrit lacks most of these prerequisites. As a consequence, transfer of knowledge between languages (English definitions to Sanskrit word senses) and between different historical domains of Sanskrit literature plays an important role in our research.

While there is, in principle, no lack of Sanskrit texts,[3] the language is nevertheless under-resourced from the perspective of NLP. First, large parts of the literature have not yet been digitized. This applies to the Sanskrit source texts and to their translations into modern languages, and complicates unsupervised knowlegde acquisition from large (parallel) corpora as, for instance, proposed by Prochasson and Fung (2011). Second, the rich morphology, especially of VS, the lack of reliable punctuation marks[4], and the phonetic phenomenon of Sandhi ("combination [of phonemes]") make linguistic analysis a hard task for NLP. Due to these features, standard token-based NLP pipelines cannot be applied to Sanskrit, as becomes apparent for a short phrase such as *saitatpaśyatītyuktvā*. This string is formed by phonetically merging the five inflected tokens *sā*, *etad*, *paśyati*, *iti*, and *uktvā* (tiny "equations" give the operative Sandhi rules):

sā [ā+e=ai]	*etad [d+p=tp]*	*paśyati [i+i=ī]*	*iti [i+u=yu]*	*uktvā*
she:N.SG.	this:A.SG.	see:3.sg., pr.	thus:ind.	say:abs.

... having said: 'She sees this' ...

Apart from the correct analysis given above, this string has at least seven further readings that are lexically valid, though semantically meaningless. Because the valid tokenization of a Sanskrit text requires a full morphological and lexical analysis, the methods described in this paper operate with fully disambiguated lemmata, instead of tokenized strings as is usually done in NLP of English.

In order to mitigate the problems introduced by size and structure of the corpus, we use bilingual information for WSD. Sanskrit has a rich history of philological research both in India and in the West. Part of this history are comprehensive Sanskrit-English dictionaries, which are also available in digital form.[5] These dictionaries provide English definitions for Sanskrit lemmata, and the definitions are ordered following lexicographic considerations. We will use the lexical definitions and their lexicographic order along with Sanskrit context words for WNSS classification in VS and CS. This setting may remind

[3] The size of Sanskrit literature can not be estimated reliably. Wujastyk (2014) considers that there exists a total of 30,000,000 Indian manuscripts, a substantial number of which may contain Sanskrit works. Several thousand Sanskrit texts have been edited and printed in the last 200 years, and a few hundred of them are available in digital form.

[4] Sanskrit texts are structured by *daṇḍa*s 'sticks' (|). These symbols indicate the end of metrical sequences, which are quite frequently not identical with sentence boundaries (Hellwig, 2016). Sentence internal structuring symbols such as commata and colons are missing completely. When applied to a Sanskrit text, the term 'sentence' refers to *daṇḍa*-delimited sequences of words in the rest of the paper.

[5] `http://www.sanskrit-lexicon.uni-koeln.de/`

of knowledge based approaches to WSD. However, it should be noted that the proposed method does not calculate the lexical overlap between the Sanskrit text and dictionary glosses for determining the best fitting word sense, as proposed by Lesk (1986) and later authors. Moreover, it does not use the graph structure of OpenCyc for WSD (Agirre and Soroa (2009) et al.). On the technical side, the paper will compare the efficiency of Maximum Entropy and of recurrent neural network models, both of which are regularly applied to WSD.

The rest of the paper is organized as follows. After an overview of related research in Section 2, Section 3 introduces the corpus and describes its semantic annotation layer. Section 4 describes how features for WSD are created, and which models are applied to the task. Section 5 compares the performance of the models and gives a short error analysis. Section 6 summarizes the paper.

2 Related Research

Although there exists a Sanskrit WordNet (Kulkarni et al., 2010), Sanskrit WSD has found little attention in research. While Kulkarni et al. (2010) and Bhingardive and Bhattacharyya (2017) concentrate on (broad) sense induction for Hindi and other modern Indian languages, Hellwig (2012) reports quantitative results only for a few ambiguous Sanskrit lemmata. Some recent studies have dealt with WSD for other classical languages such as Old English (Wunderlich et al., 2015) or Latin (Aguilar et al., 2016; Bamman and Crane, 2011). The methodology described in the last two papers (structured prediction for WSD, knowledge acquisition from parallel bilingual corpora) cannot be transferred to Sanskrit WSD, because corpora with contiguous word semantic annotations and large parallel corpora are largely missing. Similarly, diachronic WSD using word embeddings (Hamilton et al., 2016) or graphical models (Wijaya and Yeniterzi, 2011; Frermann and Lapata, 2016) cannot be applied due to the limited size of the digital Sanskrit corpus and the uncertainties in its historical stratification.

Ciaramita and Johnson (2003) introduced the task of Wordnet supersense (WNSS) classification by mapping fine-grained WordNet senses to the titles of the containing lexicographer files. The authors report accuracy rates of 52.3% on the type and 53.4% on the token level for words contained in WordNet 1.71, but not found in WordNet 1.6. This work was continued by Curran (2005), who discusses linguistic and lexicographic challenges in WNSS definition and assignment, and achieves an overall accuracy of 68% using a multi-class perceptron. Similar approaches are reported in Ciaramita and Altun (2006) and Schneider and Smith (2015). Johannsen et al. (2014) study supersense tagging for English Twitter data, using structured prediction and pretrained word embeddings. Flekova and Gurevych (2016) co-train word and supersense embeddings using the word2vec model, and construct a supersense tagger for English by feeding these embeddings along with further hand-crafted features into a multi-layer neural network. The authors obtain a classifier that performs close to the state of the art.

To sum up, the present paper is, to the best of our knowledge, the first attempt to develop a WNSS tagger for Sanskrit. Contrary to many proposed methods for WNSS of English, it relies heavily on cross-lingual information, and cannot make use of Lesk-style measures of text-gloss overlap, because texts and glosses are composed in different languages.

3 WordNet Supersenses for Sanskrit

We perform WSD with the 26 WordNet supersenses of nouns introduced in Ciaramita and Johnson (2003). WNSS are generated from the word semantic annotation layer of the Digital Corpus of Sanskrit (DCS; Hellwig, 2015). This corpus contains 4,170,064 word tokens (85,431 lexical types) with manually validated morphological and lexical annotations from all periods of Sanskrit literature, but with a strong focus on CS. 491,119 out of these 4,170,064 tokens are additionally annotated with fine-grained word semantic labels by a single annotator, using the OpenCyc (Lenat, 1995) sense inventory as starting point. Relying on the results of a single annotator is far from ideal, because there is no control of the error level, and no baseline for disagreement of human annotators can be calculated. However, as in the case of the verb-argument labelings themselves (see p. 1), other large scale annotations are not available at the

Word class	Tokens	Types
Nouns	294,506	20,307
Adjectives	67,958	3,521
Verbs	71,942	2,798
Particles, indeclinables	56,713	396

Table 1: Size of the word semantic annotation layer in the DCS: Number of lexical tokens and types with word semantic annotations, split by word classes

moment. Table 1 shows that the majority of annotated lemmata, both on token and type level, are nouns. These 294,506 sense annotated noun tokens serve for training and testing the WSD models in this paper.

Concepts not found in the original version of OpenCyc were added to the sense inventory during annotation of texts. A total of 18,804 distinct concepts were annotated in the DCS, 10,065 of which were not contained in the original OpenCyc inventory. Translations between OpenCyc concepts and WNSS were generated by first mapping OpenCyc concepts onto the English Wordnet. For finding corresponding entries, we compared the terms and the string based overlap of their definitions in OpenCyc and Wordnet. Based on this information, supersenses were retrieved from the WordNet lexicographer files. Newly created concepts, for which this mapping provides no WNSS, were labeled with the WNSS of their parent concept.[6]

While parts of the alchemical literature ($\geq$ 1300 CE) and the Bhagavadgītā (100-300 CE?) were sense annotated completely in the DCS, many semantic annotations were added to single, "philologically interesting" words; this means either to frequent words with an unusual meaning, or to rare words. The majority of these words are nouns and refer to concrete entities. The bias introduced by this annotation mode is aggravated by the text-historical composition of the DCS, because scientific (medical, alchemical) and epic texts such as the Mahābhārata (300 BCE - 500 CE?) are strongly overrepresented. The dominance of the scientific subcorpus is particularly relevant for WSD, because its vocabulary contains numerous rare technical terms denoting plants, diseases, body parts, and medical or alchemical procedures. As a consequence, senses denoting concrete entities and acts are overrepresented.

The semantic classification targets Vedic and rare nouns, and it cannot be taken for granted that these nouns show the same distribution of WNSS as frequent ones. In addition, WNSS were originally designed for modern Western texts, so that they may not cover the conceptual space of ancient Indian texts in an appropriate way. In order to understand the distribution of supersenses over noun frequency classes, we annotated three additional data sets S_{1-3} of 400 tokens each with WNSS. S_1 simulates the distribution of hapax legomena in a medium-sized corpus, and corresponds to the evaluation setting **Rare nouns** (see Sec. 5). The complete DCS is split into 20 subcorpora of approximately 200,000 tokens, respectively.[7] From each subcorpus, we randomly drew 20 tokens that are hapax legomena in their respective subcorpus. S_2 contains 400 randomly drawn hapax legomena from the Vedic layer of the DCS, and corresponds to the evaluation setting **Vedic nouns** in Sec. 5. S_3 consists of 400 randomly drawn tokens from the complete DCS, which must not be hapax legomena. S_3 is intended for simulating the composition of the training set.

The frequency distribution of supersenses in the three samples displayed in Table 2 allows several interesting insights. First, few supersenses are frequent in all three samples. When considering the nature of the Sanskrit texts and the annotation mode, high frequencies could be expected for concrete supersenses such as '*artifact*', '*person*', '*plant*', and '*substance*'. Because most instances of '*substance*' and

[6] OpenCyc is not structured in a strictly hierarchical manner. The parent concept P of a given concept C is obtained by selecting the most frequently annotated item for which a subclass relation between P and C is recorded; if such a record does not exist, P is set to the most frequently annotated item, for which an instance or member relation is recorded.

[7] We chose this size because it comes close to that of the Ṛgveda.

	S_1	S_2	S_3	N	P
person	25.25	29.84	24.26	247	26.45
act	13.13	14.4	7.35	117	11.63
comm.	4.8	17.02	2.94	88	8.25
substance	9.34	2.62	14.71	67	8.89
artifact	5.56	6.28	7.35	56	6.4
plant	10.1	1.83	3.68	52	5.2
state	3.54	5.76	2.94	40	4.08
cognition	2.53	2.09	8.82	30	4.48
attribute	3.54	2.88	2.94	29	3.12
location	3.79	1.83	5.15	29	3.59
body	3.28	2.36	3.68	27	3.11
feeling	2.53	1.57	2.21	19	2.1
animal	1.77	2.09	2.21	18	2.02
group	2.02	1.57	0.74	15	1.44
time	0.51	1.83	4.41	15	2.25
object	1.52	1.31	1.47	13	1.43
process	1.77	1.31	0.74	13	1.27
quantity	1.01	0.26	2.94	9	1.4
event	0.76	0.79	0.74	7	0.76
phen.	1.01	0.79	0	7	0.6
poss.	0.76	0.52	0.74	6	0.67
shape	0.51	0.52	0	4	0.34
food	0.25	0.52	0	3	0.26
relation	0.76	0	0	3	0.25

Table 2: Proportions of WNSS in three manually annotated samples of 400 lexical tokens; S_1: hapax legomena in 200,00 token subcorpora; S_2: hapax legomena in late Vedic texts; S_3: random tokens from the full DCS. – Rows are ordered by summed absolute frequencies (N) of supersenses in the three samples (P: proportions in the three samples). Differences to the sum of 3×400 indicate that some samples could not be labeled.

'*plant*' occur in the late medical and alchemical subcorpora, these two supersenses have low proportions in the sample from old literature (S_2). '*cognition*', '*time*', and '*quantity*' are more frequent in S_3 than in the hapax legomena samples, because they comprise generic terms such as *samaya* '(right) moment', *jñāna* 'knowledge', and number words, which are frequent, but have few synonyms.

The supersenses '*act*' and '*communication*' show a somehow opposite distribution. These supersenses are more frequent in S_1 and S_2 than in S_3 and, therefore, relevant for the main task of this paper. While '*act*' often denotes special procedures in medicine and ritual such as *mahāśānti* 'an expiatory observance and recitation', tokens annotated as '*communication*' in S_2 mostly denote special types of Vedic hymns mentioned in theoretical passages.[8] It is important to keep in mind that the ṚGVEDA, whose verb-argument annotation basically motivates this paper, comes from a different text genre than the other Vedic texts. While it also deals with the invocation of deities, it puts no emphasis on the theoretical reflection of the involved speech acts, but takes its imagery from battle, mythology, and daily life.

4 Models and Features

Classification cannot benefit from structured prediction, because the majority of annotations is attached to isolated words (Sec. 4). Therefore, we perform WSD of single words using a Maximum Entropy model

[8] The ritual handbooks called Brāhmaṇas and the Upaniṣads constitute the major part of the old layer in the DCS, from which S_2 is drawn. These texts discuss the ritual and especially ritual formulae and hymns by drawing analogies between these texts and the outer world (Hillebrandt, 1897).

(ME) and an ensemble of recurrent neural networks (RNN). This section describes the architecture of these classifiers and the features used to train them.

4.1 Maximum Entropy

We use two types of features for training the ME model (Berger et al., 1996). **Definitions** are extracted from the English glosses provided by Monier-Williams (1899). Each definition is parsed using the Stanford NLP parser (Manning et al., 2014)[9], and the syntactic root ("head") and all other nouns, adjectives, and verbs are extracted ("context"). The lexicographic definitions contain many entries of the form "a kind of plant" or "name of a warrior", where the direct syntactic dependent of the root better indicates the semantic class of the lemma than the actual root ("a kind of plant" primarily denotes a plant). These definitions are detected using the string pattern *a* (name|kind|class) of.** The dependent of the syntactic root is extracted from the parse tree of the definition, and selected as the head word of the definition. As an example, the lexicographic definition "any cry or noise" (for *ruta*) produces "cry" as head and "noise" as context word, while the definition "a particular class of gods under the Manu Tāmasa" yields the head "god" (being the direct dependent of the syntactic root "class") and the context words "particular", "class", "Manu", and "Tāmasa".

Head and context words are weighted by their lexicographic ranks in Monier-Williams (1899), because the dictionary orders the defininitions mainly, though not fully consistently by their importance.[10] Let N denote the number of definitions of a Sanskrit lemma, and r_i the 1-based rank of a single definition i, head and context words extracted from this definition are weighted with a factor w_{lex}:

$$w_{lex} = \frac{N - r_i + 1}{\sum_{j=1}^{N}(N - r_j + 1)}$$

The feature type *definitions* is generated for the target word itself, and for the two words directly preceding resp. following the target in the Sanskrit sentence.[11] Heads and context words for targets and surrounding lemmata are distinguished with prefixes.

The second type of features is the lexical context provided by the Sanskrit lemmata that surround the target word in a sentence and that are not function words (**lex_context**). Lemmata are weighted with their inverse distances to the target for ME.

The ME model is trained with limited-memory BFGS and L2 regularization of 0.5. We use the implementation from `http://www.logos.ic.i.u-tokyo.ac.jp/~tsuruoka/maxent/`.

4.2 Baseline

We use the WNSS of the first head of the first definition of a lemma as a baseline for WSD. As an example, the first definition of the lemma *viṣṇu* is given as "name of one of the principal Hindū deities". The head word of this definition according to 4.1 is "deity", which is mapped to the WNSS '*person*' as baseline prediction. If WordNet contains more than one synset containing the head word, the supersense of the first synset is chosen as prediction. If the head is not contained in WordNet, the baseline predicts the UNK tag.

4.3 Recurrent Neural Network Model

The RNN model is an extension of the architecture proposed by Tang et al. (2015) for sentiment classification. We test this kind of architecture, because we try to predict a WNSS on the basis of sentences (Sanskrit context) and phrases (English definitions). Both feature types are strictly ordered by sentence

[9] Package version 3.6.0; we use the pipeline "tokenize, ssplit, pos, lemma, parse".

[10] Much of the material contained in Monier-Williams (1899) is translated from Böhtlingk and Roth (1875), and the lexicographic order of this source influences the order in Monier-Williams (1899); see Zgusta (1988).

[11] Pre-tests with larger contexts showed no increase of accuracy.

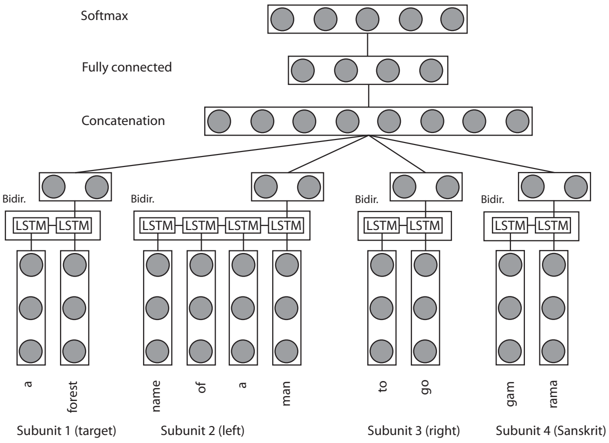

Figure 2: RNN architecture for joint training of English definitions and Sanskrit lemmata, illustrated for disambiguating the central word *vana* 'forest' in the dummy sentence *rāmo vanaṃ gacchati* 'Rāma goes into the forest'; definitions: *rāma* → "name of a man", *vana* → "a forest", *gam* → "to go".

structure ("city of god" $\neq$ "god of the city") and the lexicographic order of definitions. Choosing a recurrent architecture seemed appropriate for capturing this order. In addition, the RNN produces fixed-size numeric representations of the dictionary definitions, and thereby facilitates the transition from phrasal to lexical semantics (Hill et al., 2016).

The RNN consists of four subunits. The first three subunits receive concatenated dictionary definitions, while the fourth subunit processes the Sanskrit lemmata in the source sentence. Each subunit consists of an embedding layer of dimensions $d \times |V|$, with d denoting the embedding dimension and $|V|$ the size of the vocabulary, a bidirectional LSTM layer (Hochreiter and Schmidhuber, 1997; Graves et al., 2005) with 100 hidden units, and a dropout layer (Hinton et al., 2012) with a dropout rate of 25%. The outputs of the subunits at the last time step are concatenated and further processed with a fully connected and a softmax layer (Fig. 2). The RNN is trained with cross-entropy error as loss function, backpropagation, and a constant learning rate of 0.005 for 30 iterations.

The input for the three dictionary subunits consists of the definitions provided by Monier-Williams (1899) for the target (first subunit) and its left and right context words (second and third subunits, respectively). Definitions for each word are concatenated in their lexicographic order. Assume, for example, that Monier-Williams (1899) provides the two definitions "name of a man" and "a town" for the target word. The first subunit will receive the concatenated string "name of a man a town" in this case. The Sanskrit subunit receives the lemmata of the full sentence, where the target lemma is replaced by the UNK symbol.

The embeddings of the English words (subunits 1–3) are initialized with the pretrained vectors from the GloVe database (Pennington et al., 2014).[12] Embeddings of unknown English words are initialized with random values, and trained together with the other embeddings. In this way, words not contained in the GloVe database such as Sanskrit terms in IAST transliteration ("Śiva", "Viṣṇu"; found as "shiva" and "vishnu" in GloVe), or orthographic variants ("sun-flower") are integrated into the feature space. Embeddings of Sanskrit lemmata are pretrained using the `word2vec` tool (Mikolov et al., 2011).[13]

[12] Wikipedia 2014 + Gigaword 5 embeddings; embedding size: 50

[13] Settings: BOW, window size: 8, 5 iterations, minimal frequency of a lemma: 3, embedding size: 50. – Mixing embeddings trained with different algorithms may not be a good idea, because the choice of the training algorithm may influence how well the produced embeddings perform in specific linguistic tasks (Schnabel et al., 2015).

Rare nouns				
Classifier	Metrics	P	R	F
baseline	mi	58.46	34.82	43.64
	ma	40.43	41.71	41.06
ME	mi	76.5	76.5	76.5
	ma	53.53	53.55	53.54
RNN	mi	**77.68**	**77.68**	**77.68**
	ma	**56.19**	**55.25**	**55.71**
Vedic nouns				
Classifier	Metrics	P	R	F
baseline	mi	55.7	47.16	51.08
	ma	42.73	38.46	40.48
ME	mi	56.82	56.82	56.82
	ma	47.7	41.02	44.11
RNN	mi	**63**	**62.88**	**62.94**
	ma	**53.74**	**51.09**	**52.38**

Table 3: Results in terms of mi(cro-) and ma(cro-average) p(recision), r(ecall), and F-score; details in Table 4.

5 Experiments and Results

The models are evaluated in two settings:

Rare nouns uses the 2,809 sense annotated noun tokens whose lemmata occur less than three times in the complete DCS as test set, and the remaining sense annotated noun tokens as training set.

Vedic nouns uses the 528 sense annotated noun tokens whose lemmata occur only in the Vedic layer of the DCS as test set, and the remaining sense annotated noun tokens as training set. This setting simulates knowledge transfer from CS to VS. Note that lemmata in the test set are not required to be rare, contrary to the *rare nouns* settings.

Table 3 presents micro- and macro-averaged precision, recall, and F-scores for the two settings, while Table 4 breaks up these numbers by WNSS. Although the historical structure of the Sanskrit corpus, the annotation mode (Sec. 3), and the classifier types do not allow direct comparison with the results reported by Curran (2005) and Ciaramita and Altun (2006), Table 3 shows that ME and RNN achieve good performance, especially for *rare nouns*. Both classifiers clearly improve over the baseline. Low recall rates of the baseline indicate problems with WordNet coverage, while its low precision is caused by the interaction between high semantic ambiguity of Sanskrit nouns and lexicographic arrangement (refer to Fn. 10). For the word *aurabhra*, for example, Monier-Williams (1899) provides the definitions “a coarse woollen blanket” and “name of a physician”. Although the first meaning occurs only in indigenous monolingual dictionaries, Monier-Williams (1899) places it at first position, because it may be the etymologically older meaning (*urabhra* ‘sheep’ $\gg$ *aurabhra*). The baseline can, therefore, never access the second, correct solution.

Details in Tab. 4 demonstrate that ME and RNN have problems with nouns denoting abstract concepts.[14] While ME and RNN obtain accuracy rates of 81.7% and 83.5% for concrete nouns in the setting *rare nouns*, they only achieve 56.0% and 54.7% for abstract ones in the same setting. The higher error rate for abstract nouns can partly be explained by missing specialization of the English dictionary. In

[14] The supersenses “animal”, “artifact”, “body”, “food”, “location”, “object”, “person”, “plant”, “substance” constitute the set of concrete nouns. All other supersenses are counted as abstract.

the medical text Suśrutasaṃhitā, Cik. 11.3, for example, the term *parisaraṇa* denotes a symptom of the urinary disease called *prameha*, as a patient suffering from *prameha* "gets the habit of *parisaraṇa*." Monier-Williams (1899) glosses the hapax legomenon *parisaraṇa* as "running or moving about", which is a direct translation of the meaning "Umherlaufen" in Böhtlingk and Roth (1875). Both dictionaries were obviously not aware that the term has a medical meaning in this passage, and can best be translated as "restlessness". This translation was actually chosen as the word semantic annotation of *parisaraṇa*, and was connected with the synset "restlessness (inability to rest or relax or be still)" in OpenCyc. While the WNSS of "restlessness" is "attribute", both ME and RNN classify this occurrence of *parisaraṇa* as an "act" – a meaningful proposal given the limited amount of information available in Monier-Williams (1899) and Böhtlingk and Roth (1875).

Other misclassified instances of the WNSS "attribute" point to the problems inherent in annotating ancient languages, as in the case of the hapax legomenon *anavekṣā* mentioned in the juridicial treatise Manusmṛti (Manusmṛti, 7.111):

mohād	*rājā*	*sva-rāṣṭraṃ*	*yaḥ*	*karṣayaty*	*anavekṣayā*
folly:AB.	king:N.	own-realm:A.	who:N.	oppress:3.sg.	carelessness:I.

When a king in his folly oppresses his own realm indiscriminately, ... (Olivelle, 2005, 160)

The word was annotated with the OpenCyc concept "carelessness (the quality of not being careful or taking pains)" (WNSS: "attribute"), but labeled as "act" by ME, and as "state" by RNN. All three solutions can be justified semantically in the given textual context. While the gold annotation "attribute" fits well into the scientific character of the text, which draws a systematic picture of the ideal king, the solution "act" would highlight the voluntary negligence of royal duties. Interestingly, the semantic ambiguity is reflected, and even increased by the Sanskrit commentaries of the text. The commentary of Medhātithi seems to support a reading as an "act", because he paraphrases the term with the clause "when the king has not performed the considerations described above" (Mandlik (1886, 890); *yastu rājā pūrvoktavivekam akṛtvā ...*). On the other hand, the commentators Kullūka ("through bad teachings and lack of knowledge", *duṣṭaśiṣṭājñānena*) and Rāmacandra ("through lack of consideration", *avicāreṇa*) interpret *anavekṣā* rather as a cognitive feature or process. If their interpretation is accepted, the term should have been labeled as "cognition" in the given context. Apart from emphasizing the problem of missing adjudication (Sec. 3), this example shows the limits of semantic differentiability when interpreting ancient texts, whose languages are not spoken anymore.

ME and RNN consistently perform better for rare than for Vedic nouns. This behavior points to the problems inherent in transfering word semantic knowledge over long distances in time, and supports the conclusions reached by Sukhareva and Chiarcos (2014) for projecting parser annotations. The vast majority of training records in both settings comes from CS, so that classifiers are biased towards this form of Sanskrit. The composition of the test sets, on the other hand, shows clear differences: Out of the 2,809 words in the test set of *rare nouns*, 1,363 belong to the medical and alchemical subcorpus, 574 to the epic literature, and 247 to the poetic and narrative subcorpus, while only 118 are from the Vedic period (106 of them from the Ṛgveda). The training set of *rare words* provides plenty of data for disambiguating the WNSS of nouns from the later subcorpora, because the alchemical and the epic subcorpora are more densely annotated than other parts of the DCS (refer to page 4). These properties are not met for the setting *Vedic nouns*.

Although the ME is trained with preprocessed English definitions, the RNN produces better overall results in both experimental settings. This conclusion holds for frequent and for rare WNSS, as is evidenced by the macro-average values in Table 3 and the F-scores of rare WNSS in Table 4. We hypothesize that the ME is not able to integrate context features appropriately in several cases. In the medical passage Suśrutasaṃhitā, Nidānasthāna 9.16, for example, the word *vegāghāta* 'constipation' is correctly labeled as '*state*' by the RNN, but as '*act*' by the ME, although the head word "constipation" receives the highest coefficient of 1.529 for the label '*state*'. The misclassification of this token is caused by context features such as the Sanskrit lemma *vyāyāma* 'exertion', whose linear combination produces the final decision for '*act*'. On the other hand, performance of the RNN model drops sharply, when randomly

Rare nouns

		Baseline			ME			RNN		
WNSS	N	P	R	F	P	R	F	P	R	F
act	124	25.42	24.19	24.79	53.1	**48.39**	**50.63**	**63.24**	34.68	44.79
animal	52	**66.00**	63.46	64.71	64.44	55.77	59.79	63.79	**71.15**	**67.27**
artifact	404	60.52	34.9	44.27	**75.00**	70.54	72.7	73.85	**75.50**	**74.66**
attribute	33	24.49	**36.36**	29.27	28.21	33.33	30.56	**33.33**	30.3	**31.75**
body	44	48.78	45.45	47.06	47.73	47.73	47.73	**59.26**	**72.73**	**65.31**
cognition	14	19.23	35.71	25	**25.00**	**50.00**	**33.33**	14.29	7.14	9.52
communication	111	30.99	19.82	24.18	**73.33**	69.37	71.3	70.8	**72.07**	**71.43**
event	23	16.22	26.09	20	**50.00**	30.43	37.84	47.06	**34.78**	**40.00**
feeling	3	**30.00**	**100.00**	**46.15**	20	33.33	25	28.57	66.67	40
food	58	61.11	18.97	28.95	61.7	50	55.24	**73.17**	**51.72**	**60.61**
group	10	9.76	40	15.69	55.56	**50.00**	**52.63**	**100.00**	10	18.18
location	101	65.85	26.73	38.03	75.21	**87.13**	**80.73**	**76.15**	82.18	79.05
object	172	74.35	82.56	78.24	85.47	**85.47**	85.47	**90.74**	85.47	**88.02**
person	834	89.29	35.97	51.28	90.25	**92.09**	91.16	**93.62**	91.49	**92.54**
phenomenon	7	25	**57.14**	34.78	**50.00**	28.57	**36.36**	50	28.57	36.36
plant	196	52.27	11.73	19.17	**79.80**	80.61	**80.20**	71.25	**87.24**	78.44
possession	11	21.62	72.73	33.33	**75.00**	**81.82**	**78.26**	60	81.82	69.23
process	83	**66.67**	2.41	4.65	55.41	49.4	52.23	63.01	**55.42**	**58.97**
quantity	20	13.51	25	17.54	**52.17**	60	**55.81**	34.15	**70.00**	45.9
relation	2	0	0		0	0		0	0	
shape	1	12.5	**100.00**	22.22	0	0			0	
state	113	67.61	42.48	52.17	**68.97**	70.8	69.87	64.03	**78.76**	**70.63**
substance	378	77.99	32.8	46.18	78.61	**80.69**	**79.63**	**79.27**	79.89	79.58
time	12	29.41	41.67	34.48	**40.00**	50	44.44	38.89	**58.33**	**46.67**
Tops	3	22.22	**66.67**	33.33	**33.33**	33.33	33.33	0	0	

Vedic nouns

		Baseline			ME			RNN		
	N	P	R	F	P	R	F	P	R	F
act	51	**62.96**	33.33	43.59	62.26	**64.71**	**63.46**	57.69	58.82	58.25
animal	24	76.47	54.17	63.41	87.5	**58.33**	70	**93.33**	58.33	**71.79**
artifact	35	**66.67**	51.43	58.06	51.28	57.14	54.05	57.5	**65.71**	**61.33**
attribute	37	**57.50**	**62.16**	**59.74**	37.84	37.84	37.84	54.55	32.43	40.68
body	21	**63.16**	**57.14**	**60.00**	55.56	47.62	51.28	57.14	57.14	57.14
cognition	8	10	**12.50**	11.11	12.5	12.5	12.5	**14.29**	12.5	**13.33**
communication	32	41.94	40.62	41.27	**64.00**	50	56.14	60	**65.62**	**62.69**
event	10	18.18	**20.00**	19.05	**66.67**	20	**30.77**	14.29	10	11.76
feeling	11	43.75	**63.64**	**51.85**	20	9.09	12.5	**44.44**	36.36	40
food	3	**50.00**	33.33	**40.00**	18.18	**66.67**	28.57	16.67	66.67	26.67
group	13	16.67	**15.38**	16	**50.00**	15.38	**23.53**	33.33	7.69	12.5
location	20	**63.64**	**35.00**	**45.16**	33.33	30	31.58	46.67	35	40
object	14	46.15	42.86	44.44	26.09	42.86	32.43	**58.82**	**71.43**	**64.52**
person	163	**87.13**	53.99	66.67	70.85	86.5	77.9	76.72	**88.96**	**82.39**
phenomenon	15	42.86	40	41.38	**60.00**	20	30	42.86	**60.00**	**50.00**
plant	4	0	0		57.14	**100.00**	72.73	**60.00**	75	66.67
possession	7	60	**85.71**	**70.59**	33.33	42.86	37.5	**75.00**	42.86	54.55
process	5		0		0	0		100	**40.00**	57.14
quantity	7	0	0		50	14.29	22.22	**77.78**	**100.00**	87.5
shape	1	0	0			0			0	
state	21	61.54	**76.19**	**68.09**	50	33.33	40	**62.50**	47.62	54.05
substance	14	**77.78**	50	60.87	47.62	**71.43**	57.14	76.92	71.43	**74.07**
time	7	36.36	57.14	44.44	42.86	42.86	42.86	**55.56**	**71.43**	**62.50**
Tops	5	0	0		**100.00**	**20.00**	33.33	0	0	

Table 4: P(recision), r(ecall) and F-score for rare (upper subtable) and Vedic nouns (lower subtable). Row-wise maxima are printed bold.

initialized English and Sanskrit word embeddings are used instead of pretrained ones (macro-averaged P: 32.40, R: 34.44, F: 33.39, for the *Vedic nouns* settings; compare with Tab. 3). This finding underlines the importance of using appropriate pretrained embeddings in downstream tasks (Schnabel et al., 2015).

6 Conclusion

The paper has demonstrated that definitions from modern Western dictionaries and the lemmatized sentence context provide enough information for an efficient supersense disambiguation of rare and Vedic nouns. We would like to argue that gold information on the lemmatization level is crucial for this task, and compensates for the lack of large Sanskrit corpora to a certain degree. This indirect form of supervision is especially relevant for a morphologically rich language such as (Vedic) Sanskrit, where nouns and adjectives regularly occur in 24 case forms, and a single verbal root can produce more than 100 inflected forms. It should be noted that lemmatization not only disambiguates the Sanskrit words in the sentence context, but is equally relevant for retrieving the correct dictionary definitions of a word, which are appended to the lemma in the database of the DCS.

Future work in this area will follow two tracks. First, sense tagging was performed without using lemma information of the target word as a feature. The paper ignores the target lemma, because lemmata are by definition not useful for semantically disambiguating rare words and especially hapax legomena. It can, however, be expected that the lemma feature will clearly improve the accuracy of unrestricted Sanskrit WSD. Second, we will try to include derivational information as an additional feature in WSD of rare and Vedic nouns. Numerous Sanskrit nouns are derived from verbs or other nouns through derivational morphology, as described in Pāṇini's Aṣṭādhyāyī, or by compounding. Such derivational processes are recorded in Böhtlingk and Roth (1875) and Wackernagel and Debrunner (1954), but can also be detected using probabilistic models such as Morfessor (Creutz and Lagus, 2007). Since derivation can provide important semantic cues for the human reader, its inclusion may also improve automatic supersense disambiguation of Sanskrit nouns.

Acknowledgments

We are grateful to the anonymous reviewers for their comments. Research for this paper was supported by the Cluster of Excellence "Multimodal computing and interaction" (EXC 284, University of Saarbrücken), funded by DFG.

References

Agirre, E. and A. Soroa (2009). Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the EACL*, pp. 33–41.

Aguilar, S. T., X. Tannier, and P. Chastang (2016). Named entity recognition applied on a data base of Medieval Latin charters. The case of chartae burgundiae. In *Proceedings of the 3rd HistoInformatics Workshop*.

Bamman, D. and G. Crane (2011). Measuring historical word sense variation. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, pp. 1–10.

Berger, A., S. D. Pietra, and V. D. Pietra (1996). A maximum entropy approach to Natural Language Processing. *Computational Linguistics 22*(1), 39–71.

Bhingardive, S. and P. Bhattacharyya (2017). Word sense disambiguation using IndoWordNet. In N. S. Dash, P. Bhattacharyya, and J. D. Pawar (Eds.), *The WordNet in Indian Languages*, pp. 243–260. Singapore: Springer.

Böhtlingk, O. and R. Roth (1875). *Sanskrit-Wörterbuch*. St. Petersburg: Kaiserliche Akademie der Wissenschaften.

Che, W., T. Liu, and Y. Li (2010). Improving semantic role labeling with word sense. In *Human Language Technologies: The 2010 Annual Conference of the NAACL*, pp. 246–249.

Ciaramita, M. and Y. Altun (2006). Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of the 2006 Conference on EMNLP*, pp. 594–602.

Ciaramita, M. and M. Johnson (2003). Supersense tagging of unknown nouns in WordNet. In *Proceedings of the EMNLP*, pp. 168–175. Association for Computational Linguistics.

Creutz, M. and K. Lagus (2007, January). Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing 4*(1).

Curran, J. R. (2005). Supersense tagging of unknown nouns using semantic similarity. In *Proceedings of the 43rd Annual Meeting on ACL*, pp. 26–33.

Flekova, L. and I. Gurevych (2016). Supersense embeddings: A unified model for supersense interpretation, prediction, and utilization. In *Proceedings of the ACL*, pp. 2029–2041.

Frermann, L. and M. Lapata (2016). A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics 4*, 31–45.

Graves, A., S. Fernández, and J. Schmidhuber (2005). *Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition*, pp. 799–804. Berlin, Heidelberg: Springer.

Hamilton, W. L., J. Leskovec, and D. Jurafsky (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the ACL*, pp. 1489–1501.

Hellwig, O. (2012). ratha = "warrior" or "chariot"? Computational Approaches to Polysemy in Sanskrit. In *Proceedings of the World Sanskrit Conference 2012*.

Hellwig, O. (2015). Morphological disambiguation of Classical Sanskrit. In C. Mahlow and M. Piotrowski (Eds.), *Systems and Frameworks for Computational Morphology*, Cham, pp. 41–59. Springer.

Hellwig, O. (2016). Detecting sentence boundaries in Sanskrit texts. In *Proceedings of the COLING*, pp. 288–297.

Hettrich, H. (2007). *Materialien zu einer Kasussyntax des Ṛgveda*. Würzburg: Universität Würzburg.

Hill, F., K. Cho, A. Korhonen, and Y. Bengio (2016). Learning to understand phrases by embedding the dictionary. *Transactions of the ACL 4*, 17–30.

Hillebrandt, A. (1897). *Ritual-Litteratur. Vedische Opfer und Zauber*. Grundriss der Indo-Arischen Philologie und Altertumskunde, III. Band, 2. Heft. Strassburg: Verlag von Karl J. Trübner.

Hinton, G. E., N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Hochreiter, S. and J. Schmidhuber (1997). Long Short-Term Memory. *Neural Computation 9*(8), 1735–1780.

Johannsen, A., D. Hovy, H. M. Alonso, B. Plank, and A. Søgaard (2014). More or less supervised supersense tagging of Twitter. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics*, pp. 1–11.

Kulkarni, M., C. Dangarikar, I. Kulkarni, A. Nanda, and P. Bhattacharyya (2010). Introducing Sanskrit Wordnet. In *Proceedings on the 5th Global Wordnet Conference*, pp. 287–294.

Kulkarni, M., I. Kulkarni, C. Dangarikar, and P. Bhattacharyya (2010). Gloss in Sanskrit Wordnet. In *Sanskrit Computational Linguistics*, pp. 190–197. Springer.

Laparra, E. and G. Rigau (2013). Impar: A deterministic algorithm for implicit semantic role labelling. In *Proceedings of the ACL*, pp. 1180–1189.

Lenat, D. B. (1995). CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM 38*(11), 33–38.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, pp. 24–26.

Mandlik, V. N. (Ed.) (1886). *Mānava-Dharma Śāstra. With the commentaries of Medhātithi et al.* Bombay: Ganpat Krishnaji's Press.

Manning, C. D., M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60.

Mikolov, T., A. Deoras, D. Povey, L. Burget, and J. Černockỳ (2011). Strategies for training large scale neural network language models. In *2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 196–201.

Monier-Williams, M. (1899). *Sanskṛit-English Dictionary*. New Delhi: Munshiram Manoharlal Publishers Pvt. Ltd. (3rd edition, 1988).

Olivelle, P. (2005). *Manu's Code of Law. A Critical Edition and Translation of the Mānava-Dharmaśāstra*. Oxford: Oxford University Press.

Pennington, J., R. Socher, and C. D. Manning (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 EMNLP*, pp. 1532–1543.

Prochasson, E. and P. Fung (2011). Rare word translation extraction from aligned comparable documents. In *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies-Volume 1*, pp. 1327–1335. Association for Computational Linguistics.

Roth, M. and M. Lapata (2015). Context-aware frame-semantic role labeling. *Transactions of the Association for Computational Linguistics 3*, 449–460.

Scharfe, H. (1977). *Grammatical Literature*. A History of Indian Literature, Volume 5, Fasc. 2. Wiesbaden: Otto Harrassowitz.

Schnabel, T., I. Labutov, D. M. Mimno, and T. Joachims (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on EMNLP*, pp. 298–307.

Schneider, N. and N. A. Smith (2015). A corpus and model integrating multiword expressions and supersenses. In *Proceedings of the NAACL*.

Sukhareva, M. and C. Chiarcos (2014). Diachronic proximity vs. data sparsity in cross-lingual parser projection. A case study on Germanic. In *Proceedings of the COLING*, pp. 11–20.

Tang, D., B. Qin, and T. Liu (2015). Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on EMNLP*, pp. 1422–1432.

Wackernagel, J. and A. Debrunner (1954). *Altindische Grammatik. II, 2: Die Nominalsuffixes*. Göttingen: Vandenhoeck & Ruprecht.

Wijaya, D. T. and R. Yeniterzi (2011). Understanding semantic change of words over centuries. In *Proceedings of the 2011 International Workshop on Detecting and Exploiting Cultural diversity on the Social Web*, pp. 35–40.

Wilks, Y. (1975). A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence 6*(1), 53–74.

Witzel, M. (1995). Early Indian history: Linguistic and textual parametres. In G. Erdosy (Ed.), *The Indo-Aryans of Ancient South Asia. Language, Material Culture and Ethnicity*, Volume 1, pp. 85–125. Berlin, New York: Walter de Gruyter.

Wujastyk, D. (2014). Indian manuscripts. In *Manuscript Cultures: Mapping the Field*, pp. 159–182. Berlin: De Gruyter.

Wunderlich, M., A. Fraser, and P. Langeslag (2015). "GodWat Þæt Ic Eom God" – An exploratory investigation into word sense disambiguation in Old English. In *Proceedings of the GSCL*, pp. 39–48.

Yu, L.-C., C.-H. Wu, and J.-F. Yeh (2010). Word sense disambiguation using multiple contextual features. *Computational Linguistics and Chinese Language Processing 15*(3-4), 181–192.

Zgusta, L. (1988). Copying in lexicography: Monier-William's Sanskrit Dictionary and other cases (dvaikośyam). *Lexicographica 4*, 145–173.