silpa_nlp at SemEval-2022 Tasks 11: Transformer based NER models for Hindi and Bangla languages

Pawankumar Jawale* **Sumit Singh *** **Uma Shanker Tiwary**
Indian Institute of Information Technology, Allahabad, UP, India
{pawankumar.jawale, sumitrsch}@gmail.com
ust@iiita.ac.in

Abstract

We present Transformer based pretrained models, which are fine-tuned for Named Entity Recognition (NER) task. Our team participated in SemEval-2022 Task 11 MultiCoNER: Multilingual Complex Named Entity Recognition task for Hindi and Bangla. Result comparison of six models (mBERT, IndicBERT, MuRIL (Base), MuRIL (Large), XLM-RoBERTa (Base) and XLM-RoBERTa (Large)) has been performed. It is found that among these models MuRIL (Large) model performs better for both the Hindi and Bangla languages. Its F1-Scores for Hindi and Bangla are 0.69 and 0.59 respectively.

1 Introduction

Named Entity Recognition (NER) is one of the hot topic in natural language processing (NLP). NER is a task of identification of named entities from given sentence and their classification into predefined classes like Person, Location, Organisation, Corporation etc. For below sentence:

राम दिल्ली में गूगल में काम करता है।,
राम is Person, दिल्ली is Location and गूगल is Corporation.

The application of NER can be found in other NLP tasks such as text summarization (Toda and Kataoka, 2005), information retrieval, machine translation (Babych and Hartley, 2003), question-answering (Molla Aliod et al., 2009). The researchers have come up with many approaches for NER task such as Rule-based (Krupka and IsoQuest, 2005), feature-based Supervised approach (Liao and Veeramachaneni, 2009), Unsupervised approach and Deep learning based approach (Li et al., 2020) and Transformer based approach (Vaswani et al., 2017).

The Transformer models are good at capturing features from lengthy sentences compared to recurrent neural networks (Vaswani et al., 2017). RoBERTa model (Liu et al., 2019) performed good for NER task for rich resource languages like English.

For Hindi and Bangla languages, we applied XLM-RoBERTa, which is a multilingual version of RoBERTa pre-trained in 100 languages (including Hindi and Bangla). We also applied IndicBERT (Kakwani et al., 2020) and mBERT (Devlin et al., 2018). At last, we applied MuRIL (Khanuja et al., 2021; Sharma et al., 2022) which is specifically pre-trained in the text of 17 Indic languages and it gave better result than above models for the NER task.

This paper consists of a total of six sections apart from the introduction. Section 2 briefly defines the problem definition and task provided by organizers. Section 3 discusses the work done till now on the NER task. Section 4 mentions the dataset being used in this paper. Section 5 describes the general Transformer architecture for the NER task, along with prepossessing and post-processing. Section 6 discusses the results obtained by used models on both Hindi and Bangla languages and the error analysis. Finally, Section 7 concludes this paperwork.

2 Problem Definition

The organisers (Malmasi et al., 2022b) have arranged 13 tasks according to language. They have provided a separate dataset for each task. Each dataset is comprised of training, development and testing. Respective named entity tags were provided in the training and development dataset. Only tokens were provided in the testing dataset. Participants were required to train the models using the training and development dataset and predict NER tags on the testing dataset. We have worked on Hindi and Bangla tasks.

* Authors equally contributed to this work.

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1536 - 1542
July 14-15, 2022 ©2022 Association for Computational Linguistics

3 Related Work

Several works have been done on NER that can be categorized under two broad categories: traditional and deep learning methods.

3.1 Traditional NER approaches

In this approach, feature engineering is carried out by the researchers (Li et al., 2020). Under this category comes the rule-based, feature-based supervised learning, and unsupervised learning approaches.

In the rule-based method, the hand-crafted semantic and syntactic features are provided to recognize the entities (Krupka and IsoQuest, 2005; Aone et al., 1998). These rules-based systems can not be extended to other domains because they depend on domain-specific rules (Appelt et al., 1995).

In the feature-based supervised approach, feature engineering plays a critical role. Features such as word-level features (Liao and Veeramachaneni, 2009; Settles, 2004) and document-level features (Ravin and Wacholder, 1997; Zhu et al., 2005) are used. These features are then passed through supervised models: HMM (Eddy, 1998), Decision trees (Quinlan, 1986), SVM (Hearst et al., 1998) and CRF (Lafferty et al., 2001) for the classification in the labeled corpus.

In the unsupervised approach, the lexical patterns and statistical features are computed, which helps in the clustering (Collins and Singer, 1999). The clustering approach is applied as the data is not labeled in these cases. They extract named entities by making clusters depending on the context similarity (Nadeau et al., 2006).

3.2 Deep Learning NER approaches

As compared to the traditional NER approach, this approach does not explicitly need features. These models automatically extract the hidden features, due to which the accuracies of these models are high compared to the traditional NER approaches. This approach involves the work done using multi-level perceptrons, CNN (Wu et al., 2015), and BiLSTM (Wei et al., 2016; Lin et al., 2017). Recently the Transformer-based models have gained significant advancement in this field (Wolf et al., 2020). The Transformer-based models are good at capturing features in lengthy sentences as compared to recurrent neural networks. The Transformer (Vaswani et al., 2017) is equipped with parallel training and made up of a pair of an encoder and a decoder (to get sequence to sequence prediction). For NER task encoder is used.

In paper (Devlin et al., 2019), the authors have performed NER task on CoNLL-2003 dataset (Tjong Kim Sang and De Meulder, 2003). The authors applied both variants of BERT (Large and Base). The Large variant achieved an F1-score of 92.8 on the test set, whereas the base variant achieved F1-score of 92.4 on the test set. The authors have used BERT as word embedding and fed this to the BiLSTM. XLM-RoBERTa (Conneau et al., 2020) outperform the mBERT (Devlin et al., 2019) and XLM (CONNEAU and Lample, 2019) and show strong improvements over low-resource languages.

4 Data

The dataset (Malmasi et al., 2022a) for Hindi and Bangla contains six different NER entities, namely Location (LOC), Person (PER), Production (PROD), Group (GRP), Corporation (CORP) and Creative Work (CW). The dataset is in standard CONLL format, which uses BIO (Beginning-Inside-Outside) tagging. The dataset provided was of three types, namely training, development and testing. The training and development data contains tokens with tags, whereas testing data contains only tokens. For both Hindi and Bangla tracks, there were 15300 samples in training and 800 samples in development. In the testing dataset, for Hindi and Bangla track there were 141565 (with 933273 total tokens) and 133119

Tag	Training	Development
B-LOC	2614	131
B-PER	2418	133
B-PROD	3077	169
B-GRP	2843	148
B-CORP	2700	134
B-CW	2304	113
I-LOC	1604	77
I-PER	2836	166
I-PROD	2295	107
I-GRP	5821	297
I-CORP	2917	138
I-CW	3592	151
O	209545	10882
Total	244566	12646

Table 1: Entity distribution for Hindi track

(with 693886 total tokens) samples, respectively. Tables 1 and 2 shows the number of each entity in the training and development dataset for Hindi and Bangla, respectively.

Tag	Training	Development
B-LOC	2351	101
B-PER	2606	144
B-PROD	3188	190
B-GRP	2405	118
B-CORP	2598	127
B-CW	2157	120
I-LOC	1453	61
I-PER	3132	180
I-PROD	1964	129
I-GRP	4248	226
I-CORP	2701	122
I-CW	2844	161
O	160250	8654
Total	191897	10333

Table 2: Entity distribution for Bangla track

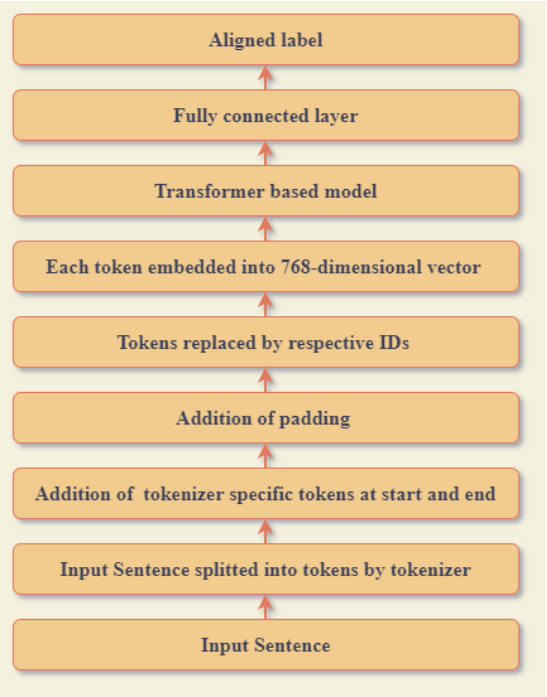

Figure 1: Generalized transformer-based model

5 Methodology

This work fine tuned 6 Transformer (Vaswani et al., 2017) based pre-trained model for the task. IndicBERT (Kakwani et al., 2020) is Albert based model which is pre-trained on 11 Indic languages, including Hindi and Bangla. We also fine-tuned XLM-RoBERTa (Base) and XLM-RoBERTa (Large) (Conneau et al., 2020), which is pre-trained on text in 100 languages. Other models are mBERT (Devlin et al., 2018) which is pre-trained on text in 104 languages and MuRIL Base and MuRIL Large (Khanuja et al., 2021) which is pre-trained on text in 17 languages with explicitly augmented monolingual text corpora with translated and transliterated document pairs. All the corpora describe above include Hindi and Bangla languages.

Figure 1 shows the architecture of this work, which is divided into 3 sections: Preprocessing, Fine tuning and Post processing.

5.1 Preprocessing

XLM-RoBERTa (Conneau et al., 2020) model and IndicBERT (Kakwani et al., 2020) uses SentencePiece tokeniser (Kudo and Richardson, 2018), which is language independent subword tokeniser and detokeniser. mBERT, MuRIL Base and MuRIL Large model uses WordPiece tokeniser (Wu et al., 2016). As all models use subword tokeniser, any token may get divided into more than one subword. Therefore an alignment of the label is required for that token. Each subword is assigned with the same label as the tokenised word. In figure 2, the token पैंजर is divided by tokeniser into two subwords: 'पै' and 'ंजर', both subwords gets B-CW as their label and the token थी। is divided by tokeniser into two subwords: 'थी' and '।', both subwords gets O as their label. Tokenised sentences are added with special tokens along with padding tokens, and thereafter, all the tokens replaced with their ID values for feeding into the models.

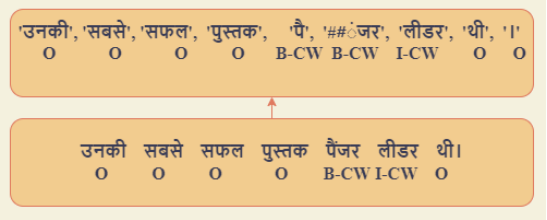

Figure 2: Label alignment

5.2 Fine tuning

Architecture in Figure 1 shows that a fully connected layer added on final output hidden vector of the model. This layer takes word embedding corresponding to each token generated by the model

(base models generate word embedding of 768 dimension and large models generate word embedding of 1024 dimension) and maps each embedding to the output layer of size (13,), which is the number of unique labels of our task. Further, we calculate loss using the cross-entropy loss function. This model is optimized with Adamw (Loshchilov and Hutter, 2019) and L2 weight decay of 0.01. It is fine-tuned with the dynamic learning rate with linear learning rate scheduler with max learning rate 4e-5, and also, batch size varies from 8 to 64 for different models subject to optimization and a dropout of 0.1 on all layers applied. Maximum token length is taken between 84 to 128 depending on the maximum length of tokenised sentences, which helps in faster training. Number of epochs for training were 30 for all the models in this experiment. We chose best model based on calculated F1-score on valid data. This work predicts the sequence of labels by the argmax of the final layer for each tokens. All models along with the fully-connected layer implemented by XXXForTokenClassification (Wolf et al., 2020), where XXX refers the corresponding model.

5.3 Post processing

After the generation of labels from the model, labels are realigned according to the detokenised sentence. This is reverse of label alignment, discussed in the Preprocessing section. The labels of all the tokens which are first tokens of their original word, are taken as generated labels.

Model	Precision (%)	Recall (%)	F1-Score (%)
M1	47.99	45.77	46.42
M2	51.05	48.08	48.97
M3	62.59	61.49	61.81
M4	**70.06**	**69.07**	**69.08**
M5	47.31	45.98	46.01
M6	51.90	47.90	49.55

Table 3: Results of each model on Hindi test data (M1: mBERT, M2: IndicBERT, M3: MuRIL Base, M4: MuRIL Large, M5: XLM-RoBERTa Base M5: XLM-RoBERTa Large)

6 Results and Analysis

Table 3 and 4 shows the macro average of Precision, Recall and F1-score of each model on testing dataset for Hindi and Bangla respectively.

Model	Precision (%)	Recall (%)	F1-Score (%)
M1	45.28	41.54	42.47
M2	43.40	37.48	38.55
M3	56.98	56.73	56.71
M4	**60.25**	**59.27**	**59.52**
M5	34.75	32.26	33.37
M6	38.74	33.07	35.45

Table 4: Results of each model on Bangla test data (M1: mBERT, M2: IndicBERT, M3: MuRIL Base, M4: MuRIL Large, M5: XLM-RoBERTa Base M5: XLM-RoBERTa Large)

Tables 6 and 7 present the Entity-wise F1 score for Hindi and Bangla testing dataset corresponding to each NER model. It has been found that MuRIL (Large) is showing the highest F1 score for each entity. It has also been observed that the F1 score for CW (Creative work) is the least among all the entities. It indicates that predicting CW is the most difficult for the model.

Sentence	अब तक का सबसे बड़ा बालिका बधू (1976 फ़िल्म)
Gold annotation	[O, O, O, O, O, B-CW, I-CW, I-CW, I-CW]
mBERT	[O, O, O, O, O, O, B-CW, I-CW, I-CW]
IndicBERT	[O, O, O, O, O, O, B-CW, I-CW, I-CW]
MuRIL	[O, O, O, O, O, B-CW, I-CW, I-CW, I-CW]
XLM-RoBERTa Large	[O, O, O, O, O, O, B-CW, I-CW, I-CW]

Table 5: Comparative analysis of a sentence from test corpus

Finally, Table 5 presents the comparative results obtained using different transformer models. Here, the MuRIL output is close to Ground annotation compared to other models.

7 Conclusion

Results show that large models are better than their corresponding base models. MuRIL (Large) model is the best among all six models described above and the second-best model is MuRIL (Base).

Entity	M1	M2	M3	M4	M5	M6
LOC	51.13	52.44	61.52	67.43	49.06	5077
PER	51.50	58.11	71.09	77.86	51.76	5589
PROD	40.57	47.78	59.54	69.01	38.93	4302
GRP	46.74	49.92	63.78	71.48	50.37	5357
CW	**38.83**	**31.64**	**51.49**	**56.95**	**33.97**	**3942**
CORP	49.77	53.95	63.46	71.78	51.96	5466
Avg.	**46.42**	**48.97**	**61.81**	**69.08**	**46.01**	**49.55**

Table 6: Entity-wise F1-score of each model for Hindi dataset
(M1: mBERT, M2: IndicBERT, M3: MuRIL Base, M4: MuRIL Large,
M5: XLM-RoBERTa Base, M6: XLM-RoBERTa Large)

Entity	M1	M2	M3	M4	M5	M6
LOC	48.91	43.38	54.56	55.93	37.99	38.03
PER	56.35	56.71	74.98	78.21	45.10	48.28
PROD	39.28	40.66	55.03	63.54	37.60	37.53
GRP	35.79	29.62	53.77	**48.03**	23.45	26.75
CW	**30.44**	**22.79**	**40.78**	48.38	**18.11**	**20.97**
CORP	44.09	38.17	61.14	63.04	38.00	41.14
Avg.	**42.47**	**38.55**	**56.71**	**59.52**	**33.37**	**35.45**

Table 7: Entity-wise F1-score of each model for Bangla dataset
(M1: mBERT, M2: IndicBERT, M3: MuRIL Base, M4: MuRIL Large,
M5: XLM-RoBERTa Base M5: XLM-RoBERTa Large)

Predicting labels corresponding to Creative Work (CW) is most challenging for all the models and predicting labels corresponding to Person (PER) is easier than predicting other labels.

References

Chinatsu Aone, Lauren Halverson, Tom Hampton, and Mila Ramos-Santacruz. 1998. Sra: Description of the ie2 system used for muc-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*.

Douglas Appelt, Jerry R Hobbs, John Bear, David Israel, Megumi Kameyama, Andrew Kehler, David Martin, Karen Myers, and Mabry Tyson. 1995. Sri international fastus systemmuc-6 test results and analysis. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Bogdan Babych and Anthony Hartley. 2003. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT Workshop on MT and Other Language Technology Tools, Improving MT through Other Language Technology Tools: Resources and Tools for Building MT*, EAMT '03, page 18, USA. Association for Computational Linguistics.

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *1999 Joint SIGDAT conference on empirical methods in natural language processing and very large corpora*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Alexis CONNEAU and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Sean R. Eddy. 1998. Profile hidden markov models. *Bioinformatics (Oxford, England)*, 14(9):755–763.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

GR Krupka and K IsoQuest. 2005. Description of the nerowl extractor system as used for muc-7. In *Proc. 7th Message Understanding Conf*, pages 21–28.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition.

Wenhui Liao and Sriharsha Veeramachaneni. 2009. A simple semi-supervised algorithm for named entity recognition. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 58–65.

Bill Yuchen Lin, Frank F Xu, Zhiyi Luo, and Kenny Zhu. 2017. Multi-channel bilstm-crf model for emerging named entity recognition in social media. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 160–165.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022a. MultiCoNER: a Large-scale Multilingual dataset for Complex Named Entity Recognition.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022b. SemEval-2022 Task 11: Multilingual Complex Named Entity Recognition (MultiCoNER). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Diego Molla Aliod, Menno Zaanen, and Daniel Smith. 2009. Named entity recognition for question answering. pages 51–58.

David Nadeau, Peter D Turney, and Stan Matwin. 2006. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. In *Conference of the Canadian society for computational studies of intelligence*, pages 266–277. Springer.

J Quinlan. 1986. Induction of decision trees. mach. learn.

Yael Ravin and Nina Wacholder. 1997. *Extracting names from natural-language text*. Citeseer.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA/BioNLP)*, pages 107–110.

Richa Sharma, Sudha Morwal, and Basant Agarwal. 2022. Named entity recognition using neural language model and crf for hindi language. *Computer Speech Language*, 74:101356.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Hiroyuki Toda and Ryoji Kataoka. 2005. A search result clustering method using informatively named entities. In *Proceedings of the 7th Annual ACM International Workshop on Web Information and Data Management*, WIDM '05, page 8186, New York, NY, USA. Association for Computing Machinery.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Qikang Wei, Tao Chen, Ruifeng Xu, Yulan He, and Lin Gui. 2016. Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks. *Database*, 2016.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Yonghui Wu, Min Jiang, Jianbo Lei, and Hua Xu. 2015. Named entity recognition in chinese clinical text using deep neural network. *Studies in health technology and informatics*, 216:624.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, ukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Jianhan Zhu, Victoria Uren, and Enrico Motta. 2005. Espotter: Adaptive named entity recognition for web browsing. In *Biennial Conference on Professional Knowledge Management/Wissensmanagement*, pages 518–529. Springer.