

# NELL's subcategories from a question answering environment

Wesley W. O. Souza<sup>1</sup>, Diorge Brognara<sup>1</sup>, João A. Leite<sup>1</sup>, Estevam R. Hruschka Jr.<sup>1</sup>

<sup>1</sup>DC-UFSCar/Federal University of Sao Carlos  
Caixa Postal 676 – 13565-905 – São Carlos-SP – Brazil

wesley\_souza@dc.ufscar.br, diorge.bs@gmail.com,

joaoaugustobr@hotmail.com, estevam@dc.ufscar.br

**Abstract.** *With advances in machine learning, natural language processing, processing speed, and amount of data storage, conversational agents are being used in applications that were not possible to perform within a few years. NELL, a machine learning agent who learns to read the web, today has a considerably large ontology and while it can be used for multiple fact queries, it is also possible to expand it further and specialize its knowledge. One of the first steps to succeed is to refine existing knowledge in NELL's knowledge base so that future communication between it and humans is as natural as possible. This work describes the results of an experiment where we investigate which machine learning algorithm performs best in the task of classifying candidate words to subcategories in the NELL knowledge base.*

**Resumo.** *Com os avanços nas pesquisas em aprendizado de máquina, processamento de língua natural, na velocidade de processamento e quantidade de armazenamento de dados, os agentes de conversação estão sendo utilizados em aplicações que anteriormente eram inviáveis. A NELL, um agente de aprendizado de máquina que aprende lendo textos da web, hoje possui uma ontologia consideravelmente grande e ao mesmo tempo que pode ser utilizada para consultas de vários fatos, também é possível expandi-la e especializar seu conhecimento. Um dos primeiros passos para se ter sucesso é refinar o conhecimento já existente na base de conhecimento da NELL, de forma que a futura comunicação entre ela e humanos seja o mais natural possível. Neste trabalho investiga-se qual algoritmo de aprendizado de máquina apresenta um melhor desempenho na tarefa de classificar palavras candidatas a subcategorias na base de conhecimento da NELL.*

## 1. Introdução

O sistema de aprendizado de máquina NELL [Carlson et al. 2010] é o primeiro sistema de aprendizado sem-fim descrito na literatura. Uma das características deste sistema é a capacidade de, autonomamente, interagir com usuários (seres humanos) de redes sociais para continuamente melhorar sua capacidade de aprendizado. Neste sentido a NELL foi também o primeiro sistema computacional a se utilizar do processo de *Conversing Learning* (tradução livre do inglês "aprender conversando") [Pedro and Hruschka Jr 2012, Pedro et al. 2013].

Uma instância da NELL teve início em 2010 e está ativa até hoje com o objetivo de aprender fatos a partir de textos da web escritos na língua inglesa. A estrutura de

sua base de conhecimento é uma ontologia, com categorias, instâncias, relações e fatos. Hoje, sua base de conhecimento passou consideravelmente dos cinquenta milhões de fatos. Entretanto, a base de conhecimento da NELL, mesmo sendo considerada grande, não possui especialização em nenhuma área de conhecimento. Fazendo uma analogia com o conhecimento humano, hoje, a NELL seria como uma adolescente que terminou o ensino médio, ela sabe o básico de cada área do conhecimento humano como história, física, química e biologia, mas não tem conhecimento avançado e detalhado sobre qualquer uma dessas.

Com isso, algumas pesquisas deram início com o objetivo de expandir a base de conhecimento da NELL e a capacidade de melhorar sua aquisição de conhecimento como em [Yang and Mitchell 2016, Dalvi et al. 2015, Pedro and Hruschka Jr 2012]. Em [Souza and Hruschka Jr 2016], foi proposta uma linguagem de conversação cognitiva (do inglês *Cognitive Conversation Language - CCL*) para agentes de conversação como *chatbots* e sistemas de perguntas e respostas (do inglês *Question Answering Systems - QAS*). O diferencial da linguagem frente as outras é a possibilidade de receber regras nas quais permitem que o agente de conversação possa acessar informações disponíveis na base de conhecimento da NELL. Assim, o agente pode responder não apenas o que está nas regras, mas pode conter respostas diferentes para vários tipos de perguntas com a mesma estrutura de pergunta e resposta. Outra vantagem da CCL é que a estrutura das regras pode ser criada, alterada ou removida não só manualmente mas também de forma automática pelo próprio agente.

No trabalho de [Souza and Hruschka Jr 2016] foi encontrada uma limitação na base de conhecimento da NELL que faz com que o agente de conversação dê uma resposta incorreta mesmo que a regra da CCL esteja certa. Certas categorias da NELL não possuem subcategorias as quais são essenciais para que um agente de conversação possa retornar uma informação válida. Um exemplo disso seria pedir para um agente de conversação, que utiliza a CCL para acessar a base de conhecimento da NELL, uma lista de filmes de comédia. Nesse caso a NELL possui uma categoria de filmes (*movie* em inglês), mas não possui informações de quais instâncias da categoria filmes são filmes de comédia. Com isso, o agente retornaria uma lista de filmes, mas não necessariamente que sejam filmes de comédia.

Com isso, o objetivo desse trabalho é investigar quais algoritmos de classificação possuem o melhor desempenho na tarefa de classificar palavras candidatas a subcategoria da base de conhecimento da NELL. Para simular um ambiente mais plausível no qual um agente de conversação se depararia, foi utilizada uma base de textos extraída a partir de fóruns de perguntas e respostas. Os resultados obtidos serão úteis para trabalhos futuros para a expansão da base de conhecimento da NELL através do Aprendizado sob Demanda (*Knowledge on Demand*), em outras palavras, um agente poderá ser capaz de identificar uma lacuna na base de conhecimento da NELL e sugerir uma alteração ou busca de novos dados para suas próximas iterações de forma automática.

Nesse artigo, a expansão da base de conhecimento da NELL pode ser vista como um estudo de engenharia de conhecimento conhecida como expansão da ontologia (*Ontology Evolution* em inglês), que consiste em manipular, modificar e atualizar uma ontologia com o tempo [Zablith et al. 2015]. Existem diferentes técnicas abordadas na literatura para expansão de ontologias [Stojanovic 2004, Noy and Klein 2004], assim como diferen-

tes objetivos e problemas que podem vir a ocorrer, porém o objetivo desse trabalho não é abordar a mudança na ontologia, mas uma abordagem que pode ser feita automaticamente através de um sistema computacional, para identificar possíveis mudanças em uma determinada ontologia: a base de conhecimento da NELL.

Na Seção 2 serão apresentados alguns trabalhos relacionados a classificação textual e abordagens para melhorar a técnica de aquisição de conhecimento da NELL. Na Seção 3 são apresentados os processos realizados desde a extração dos dados até a execução dos algoritmos de classificação. Finalizando, na Seção 4 serão discutidos os resultados obtidos e o que pode ser realizado em trabalhos futuros.

## 2. Trabalhos Relacionados

A tarefa de classificação textual é conhecida na literatura na área de Processamento de Língua Natural, uma das técnicas mais conhecidas é a *LIWC*. Os programas *LIWC* possuem duas características principais para a realização de classificação em texto, o componente de processamento e os dicionários. O componente de processamento é o programa em si, que carrega uma série de arquivos de texto e percorre palavra a palavra comparando com o dicionário de palavras, onde cada palavra do dicionário possui uma categoria. Ao final do processo, ao programa *LIWC* retorna um percentual de categorias contidas nos arquivos de textos de entrada [Tausczik and Pennebaker 2010]. Em [Bradley and Lang 1999] foi introduzida a *ANEW (Affective Norms for English Words)*, uma norma para análise de sentimentos em palavras da língua inglesa.

Outras abordagens de classificação textual surgiram nos últimos anos com diferentes objetivos e técnicas que combinam Processamento de Língua Natural e Aprendizado de Máquina. O trabalho de [Sriram et al. 2010] propõe um classificador de textos postados em microblogs utilizando *bag-of-words*. O objetivo do trabalho é classificar postagens no *Twitter* em classes genéricas como *notícia, evento, opinião e mensagens privadas*. O trabalho de [Zhang et al. 2015] aborda classificação textual propondo um algoritmo baseado em *Redes Neurais Convolucionais (ConvNets)* e compara com outras técnicas como *Deep Learning, Redes Neurais Recorrentes, bag-of-words* e *n-gramas*.

Para expandir a ontologia da base de conhecimento da NELL alguns trabalhos como [Dalvi et al. 2013, Yang and Mitchell 2016, Mohamed et al. 2011] e [Settles 2011] propõem técnicas para que essa expansão possa ser realizada de maneira automática. Em [Mohamed et al. 2011] é proposta uma técnica de expansão da ontologia adicionando novas relações, conectando grupos de instâncias conhecidas entre pares de categorias. No trabalho de [Settles 2011] é proposta uma técnica para adicionar novas subcategorias, utilizando as informações que já estão contidas na ontologia (*self-discovered*) através da análise de instâncias já conhecidas.

O trabalho de [Pedro and Hruschka Jr 2012] tem o objetivo de expandir e melhorar a base de conhecimento da NELL através de uma abordagem chamada *Conversing Learning*, onde o agente explora a base de conhecimento da NELL buscando fatos que ainda não possuem uma confiabilidade alta (acima de 90% de confiança) e que, se confirmada pode ajudar a NELL a confiar mais em outros fatos, melhorando assim a base de conhecimento de maneira considerável. Assim que o agente identifica tais fatos, o mesmo publica perguntas em redes sociais tais como *Twitter*. Com isso, seus seguidores podem responder se os fatos são ou não verdadeiros. A quantidade e forma das respostas, assim

como outros fatores, influenciam na confiança que a NELL terá sobre os fatos em questão, auxiliando assim seu processo interno de autossupervisão.

### 3. Método

O objetivo desse trabalho é avaliar se é possível desenvolver um agente computacional que aprende novas categorias para a base de conhecimento da NELL a partir da leitura de textos.

A tarefa a ser realizada é a classificação de pares de palavras candidatas a subcategoria na base de conhecimento da NELL. Foram extraídos textos de fóruns de perguntas e respostas ([www.answers.yahoo.com](http://www.answers.yahoo.com) e [www.answers.com](http://www.answers.com)), esses textos serão transformados em um conjunto de sentenças. As sentenças, por sua vez, serão processadas de forma que sejam extraídos sintagmas nominais. As palavras dos sintagmas nominais extraídas serão comparadas com a ontologia da NELL, caso uma das palavras seja identificada como uma categoria já conhecida, serão formados pares de palavras com a palavra conhecida como categoria e as outras palavras da sintagma nominal.

Um exemplo desse processo, supondo que o agente extraiu dos textos dos fóruns de perguntas e respostas a seguinte sentença: “*I need a list of good horror movies*” (Eu preciso de uma lista de bons filmes de horror - em português) ao passar a frase pelo algoritmo extrator de sintagmas nominais, tem-se as listas de palavras: *[list]* e *[good, horror, movies]*. Como a primeira lista tem apenas uma palavra, esta então será descartada, pois não é possível formar pares de palavras, já a segunda lista será utilizada para comparar cada palavra com as categorias conhecidas da ontologia da NELL, utilizou-se um lematizador em todas as palavras para reduzi-las apenas ao lema (o lema de *movies* é *movie*, por exemplo). Nesse exemplo, supomos que a NELL conheça apenas a palavra *movie* como categoria, portanto, serão formados dois pares de palavras: *[good, movie]* e *[horror, movie]*. Concluindo, cada par de palavra indica que a primeira palavra possui uma chance de ser uma subcategoria da segunda palavra que já é conhecida como categoria da ontologia em questão.

Com os pares criados, as sentenças nas quais os pares pertencem passam por outros processamentos textuais para que seja possível gerar atributos que poderão ser importantes para os algoritmos de classificação. É importante ressaltar que para a escolha desses atributos, é necessário que seja possível que um agente computacional seja capaz de gera-los automaticamente sem intervenção humana, utilizando apenas ferramentas já disponíveis.

Em seguida foi realizado um processo de rotulação dos dados. Com o conjunto de dados definitivo formado, foi possível iniciar os processos da fase de Aprendizado de Máquina. O método de teste utilizado para avaliação dos algoritmos foi a validação cruzada (*Cross Validation*). Por fim, foi realizado um teste estatístico para validar a significância do experimento.

#### 3.1. Extração

Como já descrito anteriormente, as perguntas e respostas dos fóruns foram obtidas de forma manual e com um tema específico predeterminado. Todo o texto extraído foi armazenado em um arquivo *JSON* de maneira que o assunto, a pergunta e as respostas de cada tópico pudessem ser acessados para futuras consultas ou em casos que haja a

necessidade de replicar o pré-processamento. Ao todo, foram 126 tópicos, na língua inglesa, com assuntos relacionados a atores e filmes.

Com o arquivo contendo os dados brutos foi realizada a fase de pré-processamento. Uma base de dados intermediária foi criada antes do conjunto de dados definitivo. Inicialmente foram selecionados dois atributos do arquivo de dados bruto: o assunto e corpo das questões. Como os textos foram extraídos de um ambiente não-estruturado, espera-se que algumas sentenças possam ser divididas de forma errada, podendo atrapalhar outras fases do pré-processamento e também nas fases de treinamento dos algoritmos de classificação. Após esse processo, o conjunto de dados intermediário conteve, inicialmente, uma lista de 372 sentenças.

### 3.2. Pré-processamento textual

Em seguida, cada linha dos dados intermediários passou por um pré-processamento textual aplicando tokenização, lematização, remoção de *stopwords*, extração de sintagmas nominais e análise morfossintática, utilizando nltk [Bird et al. 2008], com o objetivo de construir os atributos do conjunto de dados final. Ao final do processo, o conjunto de dados intermediário ficou com três novos atributos, como é apresentado na Figura 1, onde a coluna *sentence* armazena a sentença já extraída dos dados de origem, *tokenized* armazena as sentenças separadas por *tokens*, *tagged* armazena a classe gramatical de cada palavra após passar pela análise morfossintática e *noun-phrases* armazena as sintagmas nominais extraídas das sentenças.

	<i>sentence</i>	<i>tokenized</i>	<i>tagged</i>	<i>noun-phrases</i>
0	Bad rap actors good rap actors list?	[bad, rap, actors, good, rap, actors, list, ?]	[(bad, JJ), (rap, NN), (actor, NNS), (good, JJ)...	[[bad, rap, actor, good, rap, actor, list]]
1	list of hispanic actors?	[list, of, hispanic, actors, ?]	[[list, NN), (of, IN), (hispanic, JJ), (actor,...	[[hispanic, actor]]
2	LIST OF ACTORS AND ACTRESSES?	[list, of, actors, and, actresses, ?]	[[list, NN), (of, IN), (actor, NNS), (and, CC)...	[[list, act, ors and actresses?]]
3	Who are a and b list actors?	[who, are, a, and, b, list, actors, ?]	[(who, WP), (are, VBP), (a, DT), (and, CC), (b...	[[list, actor]]
4	List of famous black actors?	[list, of, famous, black, actors, ?]	[[list, NN), (of, IN), (famous, JJ), (black, J...	[[famous, black, actor]]

Figura 1. Cinco primeiras linhas da base de dados intermediária

Após a criação dos dados intermediários, foi possível iniciar a geração do conjunto de dados definitivo apresentado na Figura 2, onde a coluna *word* representa a palavra que será categorizada como sendo ou não uma candidata a subcategoria de uma determinada categoria (coluna *nell\_category*). Para a geração do par (*word*, *nell\_category*) foi realizada uma busca na base de conhecimento da NELL para identificar quais palavras presentes na sentenças eram uma categoria conhecida da NELL. Os pares só se formam se pelo menos uma das palavras é categoria e se as duas estão no mesmo sintagma nominal extraído. Como é possível observar na Figura 2 na coluna *sentence* é repetida a mesma sentença em todos os cinco exemplos, indicando dessa forma que é possível formar vários pares de uma mesma sentença.

Utilizando ferramentas de processamento textual disponíveis no *NLTK*, foram definidos os atributos que são mais relevantes e carregam informações que irão ajudar os algoritmos de aprendizado de máquina a terem um bom desempenho.

Os atributos *word\_tag* e *nell\_category\_tag* representam quais classes gramaticais as palavras que formam o par (*word*, *nell\_category*) pertence. O conjunto de classes gramaticais utilizado foi o *Penn Treebank* (padrão do *NLTK*). Na coluna *word\_is\_category* é

apresentado se a palavra do par também é uma categoria existente na base de conhecimento da NELL. O atributo *number\_of\_occurrences* representa a quantidade de vezes que o par ocorre na base de dados, em outras palavras, é calculado em quantas sentenças na base de dados o par se repete. O atributo *is\_category\_candidate* é o rótulo base de dados. Na Figura 2 esses dados ainda não foram rotulados.

	sentence	word	nell_category	word_tag	nell_category_tag	word_is_category	number_of_occurrences
0	Bad rap actors good rap actors list?	bad	actor	JJ	NNS	no	3
1	Bad rap actors good rap actors list?	rap	actor	NN	NNS	no	2
2	Bad rap actors good rap actors list?	good	actor	JJ	NNS	no	11
3	Bad rap actors good rap actors list?	rap	actor	NN	NNS	no	2
4	Bad rap actors good rap actors list?	list	actor	NN	NNS	no	7

**Figura 2. Cinco primeiras linhas da base de dados sem rótulo**

### 3.3. Rotulação

O processo de rotulação foi realizado manualmente. Para uma melhor concordância ao realizar o processo de rotulação, foram criadas algumas regras para considerar quais palavras podem ou não ser candidatas a subcategorias da base de conhecimento da NELL.

Definiu-se um critério de rotulação onde se restringiu alguns grupos de palavras de acordo com suas características como mostrado a seguir:

- Conjunto de caracteres que não formam palavras na língua inglesa;
- Pares que não fazem sentido, exemplo: “(year, actor)” (ator, ano);
- Características subjetivas, exemplos: “good, fast, bad” (bom, rápido, mau);
- Características que mudam com o tempo, exemplo: “young” (novo);
- Categorias mais genéricas na ontologia, exemplo: (male, actor) (ator, homem).

Com os dados prontos, o próximo passo foi realizar o experimento a partir da seleção de alguns algoritmos de classificação, essa parte será apresentada na próxima seção.

### 3.4. Análise e pré-processamento dos dados

Ao final da rotulação, o atributo classe (*is\_category\_candidate*) da base de dados, com 370 instâncias, assume dois valores possíveis: *no* e *yes* (não e sim), sendo 269 instâncias rotuladas com o valor *no* e 101 instâncias rotuladas com o valor *yes*.

Em seguida foi realizada a seleção dos atributos que farão a composição do conjunto de dados utilizado na avaliação dos algoritmos, que são: *word\_tag*, *nell\_category\_tag*, *word\_is\_category*, *number\_of\_occurrences* e o atributo classe *is\_category\_candidate*.

### 3.5. Experimento

Para a execução do experimento, foram escolhidos três modelos lineares e cinco modelos não lineares. Os modelos lineares selecionados foram o *Perceptron*, *Linear SVC* e *SGD (Stochastic Gradient Descent)*. O objetivo de utilizar esses classificadores

é identificar se os dados são ou não linearmente separáveis, indicado pelos seus valores de pontuação. Para os modelos não lineares, foram escolhidos um algoritmo baseado em SVM (*Support Vector Machine*) chamado SVC, um K-NN (*K-Nearest Neighbors*) com  $k = 7$ , um Naïve Bayes, e dois baseados em árvore de decisão, sendo um que gera uma árvore simples com o algoritmo *CART* e o método de *ensemble Random Forest*. A base de dados utilizada e o experimento estão disponíveis para acesso público em <https://github.com/MaLL-UFSCar/NELL-subcategories-QnA>.

### 3.6. Resultados e análise

Os algoritmos de aprendizado escolhidos foram avaliados utilizando uma validação cruzada com dez dobras (*10-fold cross-validation*), com a mesma separação de dados. Em cada dobra, foi considerada a acurácia do aprendiz no conjunto de teste. A Figura 3 mostra as médias da acurácia de cada método em suas dez execuções, com uma demonstração de intervalo de confiança usando o primeiro decil e o nono decil (ou 10º e 90º percentis).

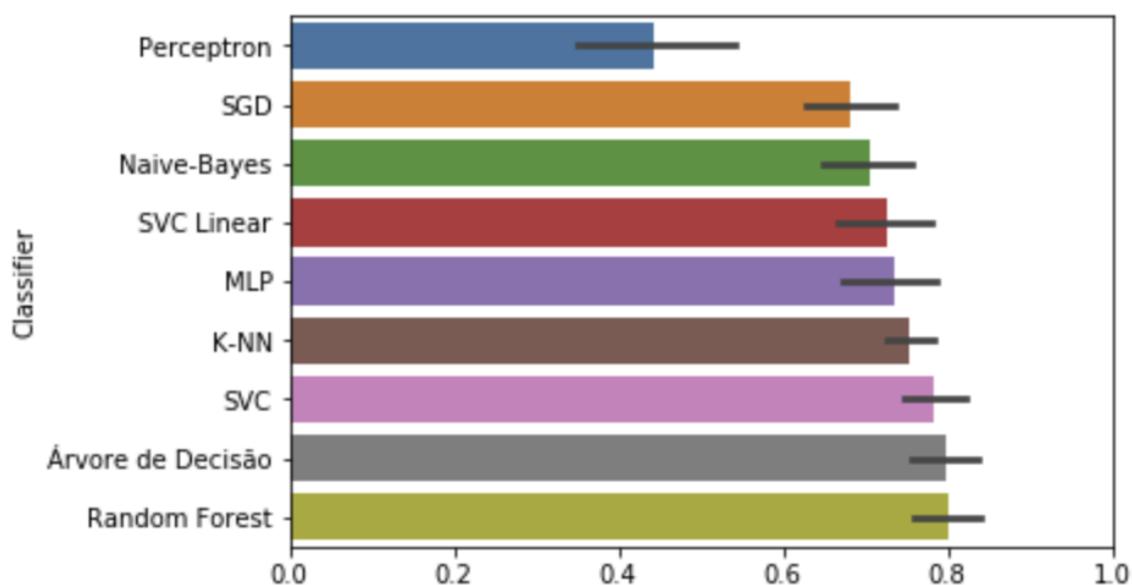


Figura 3. Média da acurácia por atributo com intervalo mínima máxima

Aplicando o método *ANOVA* aos resultados obtidos, temos que a floresta aleatória, candidato a melhor classificador, é superior aos métodos Perceptron, SGD e Naive-Bayes com significância estatística ( $p \leq 0,05$ ). Os resultados são inconclusivos para os demais aprendizes.

## 4. Conclusão

O objetivo desse trabalho foi avaliar o comportamento de alguns algoritmos de classificação em um conjunto de dados para identificação de subcategorias de uma ontologia. Com os resultados obtidos, será possível validar a possibilidade de desenvolver um agente computacional que aprende novas categorias para a base de conhecimento da NELL a partir da leitura de textos em um ambiente de perguntas e respostas.

Na fase de pré-processamento textual observou-se erros de *tokenização* no extrator de sintagmas nominais. Um exemplo disso é possível observar na linha número 2 da tabela na Figura 1, onde o texto “ACTORS AND ACTRESSES?” é dividido nos *tokens* “act” e “ors and actresses”. Essa falha não permitiu a identificação da palavra no analisador sintático, então sua TAG não foi identificada. Se no mesmo sintagma nominal houvesse uma palavra identificada como categoria da NELL, os pares com caracteres que não formam palavras na língua inglesa foram rotulados como exemplos negativos.

Em trabalhos futuros pretende-se:

- Aumentar a base de dados utilizando *webcrawlers* (algoritmos que coletam textos da web automaticamente);
- Aplicar melhorias nos algoritmos de pré-processamento textual para que haja uma tendência a diminuir a quantidade de informações incorretas no conjunto de dados;
- Utilizar o mesmo cenário em algoritmos de aprendizado semissupervisionado utilizando as respostas das perguntas como sementes.

## Agradecimentos

Agradecemos aos revisores anônimos pelas críticas construtivas. Agradecemos à CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelas bolsas de estudo e pesquisa que deram suporte para que este trabalho e outros relacionadas fossem possíveis. Também somos gratos aos nossos amigos do MaLL (*Machine Learning Lab*) pelo apoio.

## Referências

- Bird, S., Klein, E., Loper, E., and Baldridge, J. (2008). Multidisciplinary instruction with the natural language toolkit. In *Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics*, pages 62–70. Association for Computational Linguistics.
- Bradley, M. M. and Lang, P. J. (1999). Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Citeseer.
- Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr, E. R., and Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. In *AAAI*, volume 5, page 3.
- Dalvi, B., Cohen, W. W., and Callan, J. (2013). Classifying entities into an incomplete ontology. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 31–36. ACM.
- Dalvi, B., Minkov, E., Talukdar, P. P., and Cohen, W. W. (2015). Automatic gloss finding for a knowledge base using ontological constraints. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*.
- Mohamed, T., Hruschka, E., and Mitchell, T. (2011). Discovering relations between noun categories. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1447–1455, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Noy, N. F. and Klein, M. (2004). Ontology evolution: Not the same as schema evolution. *Knowledge and information systems*, 6(4):428–440.

- Pedro, S. D., Appel, A. P., and Hruschka Jr, E. R. (2013). Autonomously reviewing and validating the knowledge base of a never-ending learning system. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1195–1204. ACM.
- Pedro, S. D. and Hruschka Jr, E. R. (2012). Conversing learning: Active learning and active social interaction for human supervision in never-ending learning systems. In *Ibero-American Conference on Artificial Intelligence*, pages 231–240. Springer.
- Settles, B. (2011). Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Souza, W. W. O. and Hruschka Jr, E. R. (2016). Cognitive conversation language-ccl. In *International Conference on Intelligent Systems Design and Applications*, pages 309–318. Springer.
- Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M. (2010). Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 841–842. ACM.
- Stojanovic, L. (2004). Methods and tools for ontology evolution.
- Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.
- Yang, B. and Mitchell, T. (2016). Joint extraction of events and entities within a document context. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.
- Zablith, F., Antoniou, G., d’Aquin, M., Flouris, G., Kondylakis, H., Motta, E., Plexousakis, D., and Sabou, M. (2015). Ontology evolution: a process-centric survey. *The knowledge engineering review*, 30(1):45–75.
- Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level convolutional networks for text classification. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 649–657. Curran Associates, Inc.