의미 부착이 없는 데이터로부터의 학습을 통한 의미 중의성 해소

박성배[0] 장병탁 김영택

서 울 대 학 교 컴 퓨 터 공 학 부

sbpark@nova.snu.ac.kr {btzhang,ytkim}@cse.snu.ac.kr

Word Sense Disambiguation From Unlabelled Data

Seong-Bae Park[0] Byoung-Tak Zhang Yung Taek Kim

School of Computer Science and Engineering, Seoul National University

요 약

의미 모호성 해소는 문맥상의 한 단어의 올바른 의미를 밝히는 것으로, 대부분의 자연언어처리 응용에서 가장 중요한 문제 중 하나이다. 말뭉치로부터 얻어진 예제로부터 의미 모호성 해소 방법을 학습하기 위해서는 답이 알려져 있는 대량의 학습 예제가 필요하지만, 답이 알려져 있는 예제를 구하는 일은 사람의 간섭을 필요로 하므로 매우 비싼 작업이다. 본 논문에서는 답이 알려져 있는 학습 예제로 어느 정도 학습한 후, 답이 알려져 있지 않은 예제로 학습을 보충하는 방법을 통해 사람의 간섭을 최소화하였다. 결정트리 학습을 통한 한국어 명사에 대한 의미 결정 실험 결과, 본 논문에서 제안한 방법은 가장 많은 분포를 보이는 의미를 선택하는 경우보다 평균적으로 33.6%의 성능 향상을 보이며, 이는 전체 학습 예제의 답이 모두 알려져 있는 경우와 거의 비슷한 결과이다. 따라서, 한국어와 같이 신뢰할 만한 의미 부착 말뭉치가 없는 경우에 본 논문에서 제시된 방법은 매우 효율적이다.

1. 서론

의미 모호성 해소(WSD)는 문맥상의 한 단어의 올바른 의미를 밝히는 것으로, 대부분의 자연언어처리 응용에서 가장 중요한 문제 중 하나이다. 대규모의 말뭉치와 다양한 기계학습 기법이 사용 가능해 짐에 따라, 이 문제에 대해 말뭉치에 기반한 연구가 가능하게 되었다[2,5,6,9]. WSD 에 대한 말뭉치 기반의 연구를 위해서는, 의미 부착 말뭉치나 2 개 국어 이상의 정렬 말뭉치가 필요하다.

하지만, 한국어에 대한 WSD 에서는 다음과 같은 것들을 고려하여야 한다.

1. 한국어에는 신뢰할 수 있고 사용가능한 의미 부착 말뭉치가 없다.
2. 한국어 단어의 90% 이상이 의미 모호성을 지닌다.
3. 대규모 말뭉치에 의미를 부착하는 일은 매우 비싼 작업이다.

조정미와 김길창은 품사 부착 말뭉치만을 이용해서 한국어 동사의 의미를 판별하였다[3]. 그들은 의미 부착 말뭉치가 없어서 생길 수 있는 데이터 부족 문제를 해결하기 위해 명사 분포와 동사 분포를 사용하였다. 하지만, 이 분포가 목적어-동사 관계만을 이용하였기 때문에, 응용이 타동사로만 제한되었다. Atsushi 는 의미 결정 학습에 소수의 예제만을 사용하기 위해서 선택 샘플링 기법을 사용하였다[2]. 그들은 확실성이 가장 떨어지는 예제를 선택하기 위해서 학습 유용도 함수(TUF)를 정의한 후, 학습의 단 반복마다 확실성이 가장 낮은 예제를 예제 데이터베이스에 저장하였다. 그러나, 학습 유용도를 *k*NN 과 비슷한 방법으로 구현하였기 때문에 각 반복마다 단어 속성 벡터들 사이의 유사도를 계산하여야 하는 문제를 보였다. 반면에, Hwee 와 Lee 는 단어 의미 중의성 해소에서 여러가지 지식원을 함께 사용함으로써 상당히 높은 정확성을 얻었다[5]. 문서 분류 분야에서는, Liere 와 Tadepalli 는 다수결을 이용한 능동 *학습(Active Learning)*이 적은 수의 학습 예제로도 잘 수행됨을 보였다[6].

본 논문에서는 위원회(committee)에 의한 선택 샘플링 알고리듬을 이용하여 단어의 의미 모호성을 해소하는 새로운 방법을 제시한다. 예제가 하나 주어졌을 때, 그 예제를 학습할지 말지를 여러 분류기(classifier)들의 가중 다수결로 결정한다. 초기에 어느 정도 학습된 분류기에 답이 알려져 있지 않은 학습 예제를 사용하여 보충함으로써 사람의 간섭을 최소화하였다. 따라서, 이 방법은 적은 수의 학습 예제만으로도 의미 모호성 문제를 효과적으로 해결하므로, 의미 부착 말뭉치가 없는 언어에서의 의미 모호성 해소에 대한 가능성을 제시한다. 또한, 제시된 방법은 의미 모호성 문제뿐만 아니라 다른 종류의 분류 문제에도 적용될 수 있다.

2. 단어의 의미 모호성 문제

Atsushi 는 한국어와 비슷한 특징을 갖는 일본어에 대한 단어 의미 중의성 해소에서 격표지가 중요한 역할을 함을 실험적으로 보였다[1]. 그리고, 영어의 단어 의미성 해소에서는 이웃 단어의 품사나 형태소와 같은 여러가지 지식원이 사용되었다[5].

한국어의 특성을 표현하기 위해서, 다음 사항을 고려하여야 한다.

- 한국어는 부분 자유 어순 언어이어서, 이웃 단어에 대한 정보가 무의미할 수 있다.
- 한국어에서는 생략이 자주 일어나서, 격 의미를 포함하는 대규모의 예제를 모으기가 어렵다.

이런 점들을 고려하여, 한국어 명사에 대한 의미 모호성 해소를 위해 여덟 개의 속성을 선택하였다(표 1). 이들 중 셋(*PARENT, NMODWORD, ADNWORD*)은 그 값으로

형태소를 취하고, 하나(*GFUNC*)는 문법 기능[1]의 11 가지 값을 가지고, 나머지는 *참* 혹은 *거짓*을 값으로 가진다.

의미 모호성 해소를 위한 분류기로 결정 트리를 사용하였다. 속성 값들이 이산적이고 몇 속성에 대해서는 값 손실이 있을 수 있으므로, 결정 트리는 이 문제에 대한 좋은 분류기로 믿어진다. 본 논문에서는 Quinlan 의 C4.5 release 8 을 결정 트리 학습기로 사용하였다[7].

C4.5 를 위의 속성 벡터에 직접 적용하면, 데이터 부족 문제가 심각하게 나타날 수 있다. 그러므로, C4.5 는 속성 *PARENT*, *ADNWORD*, *NMODWORD* 에 대해서 단순 형태소 비교가 아니라 단어 사이의 유사도를 계산하여 매칭을 하도록 변경되었다.

속성	내용
GFUNC	*W* 의 문법 기능
PARENT	*W* 의 수식을 받는 단어
SUBJECT	*PARENT* 가 주어를 갖는지의 여부
OBJECT	*PARENT* 가 목적어를 갖는지의 여부
NMODWORD	*W* 를 수식하는 명사어구
ADNWORD	*W* 를 수식하는 관형절
ADNSUBJ	*ADNWORD* 가 주어를 갖는지의 여부
ADNOBJ	*ADNWORD* 가 목적어를 갖는지의 여부

표 1 : 의미 모호성이 있는 한국어 명사 *w* 의 의미를 결정하는 데 사용되는 속성.

3. 위원회에 의한 선택 샘플링

학습에 필요한 예제의 수를 줄이고, 답이 주어지지 않은 예제를 활용하는 방법으로 선택 샘플링(selective sampling)을 사용한다. 그림 1 은 의미 분류기를 학습하는 의사 코드이다. 이 알고리듬은 분포 W 가 분류기에 대한 것임을 제외하면 **AdaBoost.M1**[4]과 매우 비슷하다. 답이 주어져 있는 초기 학습 예제 L 은 재샘플링을 통해 임의로 M 개로 나누어 진 후, M 개의 분류기에 의해 각각 학습된다. 답이 주어지지 않은 각 예제 x_t 에 대해서, 각 분류기 $C_j(1 \le j \le M)$는 각각 x_t의 의미를 결정한다. 분류기에 대한 가중치 분포 W_t를 가지고 각 C_j의 결정으로부터 x_t 에 대한 추정값인 s_t 가 다수결로 결정된다. 그리고, 확실도 α_t를 계산한다. ε_t가 일종의 오류률이기 때문에, ε_t가 작을수록 α_t 가 크다. 그러므로, 한 분류기의 영향력은 학습을 하는 동안 오류를 적게 만들수록 증가하고, 많이 만들수록 감소한다.

만약 α_t 가 확실도 한계치 θ보다 크면 분류기는 이 예제 x_t 를 학습할 필요가 없다. 이 경우에는 각 분류기의 가중치를 재조정하기만 한다. α_t 가 확실도 θ보다 작다면 분류기는 x_t 의 의미가 s_t 라는 가정하에 x_t 를 학습한다. θ 의 값은 실험적으로 결정된다.

3.1 의미 결정

그림 1 의 두번째 단계에서 의미가 모호한 단어 w 의 속성 벡터인 입력 x 의 의미를 결정한다. x 의 의미 s_t 는 가중 다수결로 결정된다.

$$s_t = \arg\max_{s \in SS(w)} \sum_{j:C_j(x)=s} W(j)$$

Given unlabeled example set $D = \{x_1, \ldots, x_T\}$
And labeled example set L
Initialize $W_1(j) = 1 / M$
Resample $S_j^{(0)}$ from L for each classifier C_j where $|S_j^{(0)}|=|L|$ as done in Bagging.
Train each base classifier $C_j(1 \le j \le M)$ from $S_j^{(0)}$.
For t = 1, ..., T:

1. Determine the sense of $x_t \in D$ from each C_j.
   $S = \langle s_1, \ldots, s_M \rangle$
2. Find the most likely sense s_t from S using distribution W.
3. Set $\alpha_t = (1-\varepsilon_t) / \varepsilon_t$ where
   $$\varepsilon_t = \frac{\text{No. of classifiers whose output is not } s_t}{M}.$$
4. **If** α_t is larger than a certainty threshold θ, **then** update W_t.
   $$W_{t+1}(j) = \frac{W_t(j)}{Z_t} \times \begin{cases} \alpha_t & \text{if } s_j = s_t \\ 1 & \text{otherwise} \end{cases}$$
   where Z_t is a normalization constant.
5. **Otherwise** every classifier C_j is restructured from $S_j^{(t)}$.
   $$S_j^{(t)} = S_j^{(t-1)} + (x_t, s_t)$$

Output the final classifier:
$$C(x) = \arg\max_{s \in SS(x)} \sum_{t:C_t(x)=s} W_T(j)$$

그림 1 : 단어 의미 모호성 해소를 위한 선택 샘플링 알고리듬.

여기서 $SS(w)$는 w 의 가능한 의미로 이루어진 집합이다. 각 반복 t 에서는 가중치 분포 W_t 가 사용되고, 최종적으로는 W_T가 사용된다.

3.2 단어 유사도

세 개의 속성이 형태소를 그 값으로 취하기 때문에, 데이터 부족 문제가 발생할 수 있다. 이 문제는 시소러스나 단어 클래스 등을 이용하여 극복할 수 있다. 하지만, 한국어에는 신뢰할 만한 시소러스가 존재하지 않을 뿐더러 문법기능 부착 말뭉치도 없으므로 통계적인 방법으로 단어 클래스를 만들기도 어렵다.

본 논문에서는, 두 한국어 단어 사이의 유사도를 한영 사전과 영어 시소러스인 *WordNet* 을 이용하여 결정한다. 두 단어 W_1 과 W_2 사이의 유사도 $Sim(W_1, W_2)$는 두 한국어 단어의 영어 대역어 사이의 평균 유사도로 결정된다. 두 단어 W_1 과 W_2 는 $Sim(W_1, W_2)$이 실험적으로 정해진 한계치보다 크면 매치된다.

4. 실험

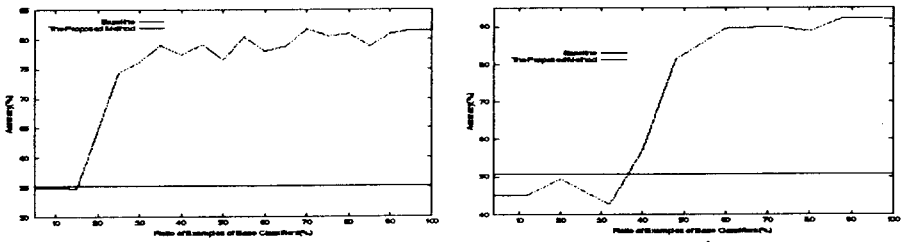
4.1 데이터

KORTERM 에서 분포한 KAIST 말뭉치를 이용하여 실험하였다. KAIST 말뭉치가 천만어절로 구성되었지만, 본 실험에서는 신문 기사를 제외하고 백만 어절 크기의 말뭉치만을 고려하였다. 표 3 은 의미 모호성이 있는 한국어 명사와 각 명사가 가질 수 있는 의미들을 나타내고, 표 2 는 실험에서 사용된 학습 예제의 수를 나타낸다. 표 3 의 *비율* 열은 각 단어가 말뭉치에서 그 의미로 사용된 비율을 나타낸다. 그러므로, 우리는 이 값을 각 단어에 대한 모호성 해소의 기준점으로 삼을 수 있다.

단어	학습 예제의 수
배	876
분	796

표 2 : 실험에 사용된 학습 예제의 수

[1] 본 논문에서 사용된 문법 기능은 서울대학교에서 개발 중인 한국어 파서에서 사용되는 문법 기능 11 가지이다.



(a) 배 (b) 분

그림 2 : 한국어 명사에 대한 의미 결정 실험 결과. X 축은 초기에 분류기가 학습한 답이 있는 예제의 전체 예제 수에 대한 비율이고 Y 축은 의미 결정의 정확도이다. 답이 있는 예제의 비율은 매 5%정도마다 측정되었으며, 정확도는 각 측정마다 10-fold cross validation 으로 측정한 평균값이다.

단어	의미 수	의미	비율
배	4	Pear	6.2%
		Ship	**55.2%**
		Times	13.7%
		Stomach	24.9%
분	3	Person	46.2%
		Minute	**50.8%**
		Indignation	3.0%

표 3: 실험에 사용된 한국어 명사의 다양한 의미

4.2 실험 결과

L 개의 의미를 갖는 명사에 대해서, $L+1$ 개의 분류기를 사용하였다. 표 4 는 표 3 에 있는 명사에 대한 의미 모호성 실험의 10-fold cross validation 결과를 보인다. 표 4 의 의미 결정 정확도는 정확도가 가장 높은 순간에 측정된 것이다. 이 실험 결과에 따르면, 답이 주어진 학습 예제를 조금만 사용하고 답이 주어지지 않은 예제로 보충한 방법이 전체 예제로 답이 있는 것을 쓰는 방법만큼 잘 수행됨을 알 수 있다. 또한, 제안된 방법은 기준점보다 평균적으로 33.9%의 성능 향상을 보였다.

그림 2 는 각 분류기에 대한 답이 주어진 초기 학습 예제의 수가 전체 성능에 어떤 영향을 미치는지를 보인다. 그림 2 의 X-축은 전체 학습 예제에 대해 처음에 분류기가 학습한 답이 주어진 예제의 비율이다. 그림 2(a)는 명사 '배'에 대한 정확도를 보인다. 초기 학습 예제의 수가 증가하면, 정확도도 증가하지만, 약 35% 정도에서 거의 평형을 이룬다. 따라서, 약 35% 정도의 답이 주어진 예제만 사용하여 학습을 하면, 전체적으로 높은 성능을 보인다고 할 수 있다. 또한, 명사 '분'에 대해서도 비슷한 현상을 보인다.

단어	제안된 방법	하나의 C4.5	기준점
배	81.5 ± 7.7%	**82.3 ± 5.9%**	55.2%
분	92.3 ± 7.7%	**94.3 ± 5.7%**	50.8%
평균	86.9%	**88.3%**	53.0%

표 2: 한국어 명사에 대한 실험 결과

5. 결론

본 논문에서는 위원회에 의한 선택 샘플링으로 의미 모호성을 해소하는 방법을 제시하였다. 초기에 어느 정도 학습된 분류기에 답이 주어지지 않은 예제를 보충함으로써 성능 향상을 얻었고, 예제를 보충할 때 분류기의 위원회의 가중 다수결을 사용함으로써 필요한 학습 예제의 수를 줄였다.

본 논문에서는 분류기로 결정트리를 사용하였기 때문에, 학습시 매 반복마다 결정트리를 재구성하여야 하는 부담이 있다. 하지만, Utgoff 등은 빠른 시간에 결정트리를 재구성할 수 있음을 보였다[8]. 따라서, 결정트리의 재구성은 부담이 되지 않는다. 하지만, 다른 방법을 분류기로 사용하였을 때, 분류기의 재구성을 어떻게 할 것인지에 대해 더 연구하여야 한다.

감사의 글

본 연구는 정보통신부 대학기초연구지원사업인 "지능형 인터넷 정보서비스를 위한 대규모 텍스트 분류 및 검색 기술 개발"(과제번호 98-199)에 의하여 일부 지원되었음.

6. 참고 문헌

[1] Atsushi F., Kentaro I., Takenobu T., and Hozumi T. "To What Extent Does Case Contribute to Verb Sense Disambiguation?" In *Proceedings of COLING-96*, pp. 59-64, 1996.

[2] Atsushi F, Kentaro I., Takenobu T., and Hozumi T., "Selective Sampling of Effective Example Sentence Sets for Word Sense Disambiguation," *Computational Linguistics*, Vol. 24, No. 4, pp. 573-597, 1998.

[3] Cho J. and Kim G., "Korean Verb Sense Disambiguation Using Distributional Information From Corpora," In *Proceedings of Natural Language Processing Pacific Rim Symposium (NLPRS)*, pp. 691-696, 1995.

[4] Freund Y. and Schapire R., "Experiments with a New Boosting Algorithm," In *Proceedings of the Thirteenth International Conference on Machine Learning*, pp. 148-156, 1996.

[5] Hwee T. and Lee H., "Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exmaplar-Based Approach," In *Proceedings of the 34th Annual Meeting of the ACL*, pp. 40-47, 1996.

[6] Liere R. and Tadepalli P., "Active Learning with Committees for Text Categorization," In *Proceedings of AAAI-97*, pp. 591-596, 1997.

[7] Quinlan R., *C4.5: Programs For Machine Learning*, Morgan Kaufmann Publishers, 1993.

[8] Utgoff P., Berkman N., and Clouse J., "Decision Tree Induction Based on Efficient Tree Restructuring," *Machine Learning*, Vol. 29, pp. 5-44, 1997.

[9] Yarowsky, D. "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods," In *Proceedings of the 33rd Annual Meeting of the ACL*, pp. 189--196, 1995.