Word sense induction using independent component analysis

Petr Šimon
Institute of Linguistics
Academia Sinica, Taiwan
sim@klubko.net

Jia-Fei Hong
Graduate Institute of Linguistics
NTU, Taiwan
jiafei@gate.sinica.edu.tw

Abstract

This paper explores the possibilities of using independent component analysis (ICA) for features extraction that could be applied to word sense induction. Two different methods for using the features derived by ICA are introduced and results evaluated. Our goal in this paper is to observe whether ICA based feature vectors can be efficiently used for word context encoding and subsequently for clustering. We show that it is possible, further research is, however, necessary to ascertain more reliable results.

1 Introduction

Word senses are known to be difficult to discriminate and even though discrete definitions are usually sufficient for humans, they might pose problems for computer systems. Word sense induction is a task in which we don't know the word sense as opposed to more popular word sense disambiguation.

Word sense can be analyzed by observing behaviour of words in text. In other words, syntagmatic and paradigmatic characteristics of a word give us enough information to describe all it's senses, given that all it's senses appear in the text.

Based on this assumption, many techniques for word sense induction have been proposed. All are based on word co-occurrence statistics. There are two strategies for creating the vectors that encode each word: global encoding strategy, which encodes co-occurrence of word types with other word types and local encoding strategy which encodes co-occurrence of word tokens with word types. The global encoding strategy is more popular, because it provides more information and does not suffer from data sparseness and most of the research has focused on sense analysis of words of different forms, i.e. on phenomena like synonymy etc. However, by encoding word types, we naturally merge all the possible sense distinctions hidden in word's context, i.e. context of a token. For more details cf. (3; 11; 10).

Problem of high dimensionality that would be computationally restricting, us usually solved by one of several methods: principal component analysis (PCA), singular value decomposition (SVD) and random projection (RP) and latent semantic analysis, also known as latent semantic indexing is a special application of dimensionality reduction where both SVD and PCA can be used. See (1; 2) for overview and critical analysis.

The classical approach to word context analysis is a vector space model, which uses simple the whole co-occurrence vectors when measuring word similarity. This approach also suffers from a problem similar to data sparseness, i.e. the similarity of words is based on word forms and therefore fails in case where synonym rather than similar word form is used in the vector encoding (11; 10).

Major problem with the classical simple vector space model approach is the superficial nature the information provided by mere co-occurrence frequency, which can only account for seen variables. One of the most popular approaches to word context analysis, latent semantic analysis (LSA), can improve this limitation, by creating a latent semantic space using SVD performed on word by document matrix. Frequency of occurrence of each word in a document represents each entry w_{ij} in the matrix, thus, the whole document serves as a context. Document is, naturally, some sort of meaningful portion of text. SVD then decomposes the original matrix into three matrices: word by concept matrix, concept by concept matrix and concept by document matrix. The results produced by LSA are, however, difficult to understand for humans (9), i.e. there is no way of explaining their meaning.

2 ICA

Independent component analysis (ICA) (7) is a statistical method that takes into account high order statistical dependencies. It can be compared to PCA in the sense that both are related to factor analysis, but PCA uses only second-order statistics, assuming Gaussian distribution, while ICA can only be performed on non-Gaussian data (6). Comparison with SVD is provided by (12) on word context analysis task.

ICA is capable of finding emergent linguistic knowledge without predefined categories as shown in (4; 5) and others.

As a method for feature extraction/dimensionality reduction it provides results that are approachable by humans reader. Major advantage of ICA is that it looks for factors that are statistically independent, therefore it is able find important representation for multivariate data.

ICA can be defined in a matrix form as $\mathbf{x} = \mathbf{A}\mathbf{s}$ where $\mathbf{s} = (s_1, s_2, ..., s_n)^T$ represents the independent variables, components, and the original data is represented by $\mathbf{x} = (x_1, x_2, ..., x_n)^T$, which can be decomposed into $\mathrm{s} \times \mathbf{A}$, where $\mathbf{A}$ is a $n \times n$ square mixing-matrix.

Both the mixing-matrix $\mathbf{A}$ and independent components s are learning by unsupervised process from the observed data $\mathbf{x}$. For more rigorous explanation see (7).

We have used FastICA algorithm as implemented in R language[1].

3 Data collection

The context matrix has been constructed from words from Sinica Corpus of a frequency higher than 150. This restriction yielded 5969 word types. We have chosen this limited lexicon to lower the complexity of the task.

The whole corpus was stripped from everything but all words whose word class tag started with N, V, A or D. This means that our data consisted of nouns (N, including pronouns), verbs (V), adjectives (A) and adverbs (D) [2].

Then we collected co-occurrence statistics for all words from window of 4 preceding and 4 following words, but only if these were within a sentence. We defined sentence simply as a string of words delimited by ideographic full-stop, comma, exclamation and question mark (。, ，, ！ and ？). In case of context being shorter than 4 words, the remaining slots were substituted by zero indicating no data available.

We have normalized the data by taking log of each data point a_{ij} in context matrix. Since this is a sparse matrix and lot of data points are zero, one has been added to each data point.

After the extraction of the independent components, we have encoded contexts of word tokens for each word type selected for analysis using these independent components. Thus we are able to provide reliable encoding for words, which is based on global properties. Note that there is no need to pursue orthogonality of different word types that are sometimes required in the context encoding. The

[1] http://www.stats.ox.ac.uk/ marchini/software.html

[2] For complete list see: http://wordsketch.ling.sinica.edu.tw/gigaword_pos_tags.html

similarities between different word types are based on the strength of independent components for each word type and therefore much better results of similarity measure can be expected than one would get from binary random encoding as introduced in (8).

We could experiment with several strategies to context matrix construction: different word classes in the context and different sizes of feature vectors. Context in our experiments is defined by four words that precede and four words that follow each keyword. Then we study the feature similarities across different words. To aid the analysis, a hierarchical clustering is used to determine closeness of relation among feature vectors of specified dimension. This step is to find most reliable feature vector dimension for subsequent experiments. As mentioned before, the features can be traced back and their nature determined, i.e. they can be labelled.

Due to the time constraints, we have predetermined feature vector size beforehand. We've extracted 100 and 1000 independent components and used them in two separate experiments.

Having determined the size of feature vectors, we use original word contexts for each word token and encode the context using these vectors. That means that each word in the context of particular keyword is replaced by it's respective feature vector, a vector of quantified relations to each of the independent components that has been extracted by ICA from the global co-occurrence matrix.

We than use maximum-linkage hierarchical clustering to find related words and based on the features present in the vectors we determine their characteristics that will provide clues to their word senses.

4 Results

We ran two experiments, one with 100 independent components and the second with 1000 components. For the experiment we have manually selected 9 words, which we expected to be easier to analyze. We have, however, failed to find in Chinese word that would allow for such obvious sense distinctions as English *plant*,*palm*, *bank* etc. Such words are typically used in word sense related task to test the new algorithms. The failure to find words that would have similarly clear-cut sense distinctions, might have influenced our initial results. The words we have selected are (number in bracket indicates the number of senses according to Chinese Wordnet)[3]: 山頭 (3), 名牌(2), 犯規(2), 約談(2), 措辭(2), 堅硬(2), 富有(2), 屬下(2), 天氣(3).

[3] http://cwn.ling.sinica.edu.tw

4.1 Independent components

When ICA algorithm retrieves the specified number of independent components, each of them can be labelled by creating a descending list of those words that are most responsive for each of the components (5; 4). Only the most responsive word could be assigned to each of the components as a label, but this way we would not be able to determine characteristics of the components with sufficient clarity. As we will see, even listing several items from the top of the list of the most responsive words, won't always provide clear explanation of the nature of the component in question. This is due to the fact that the independent components are not yet very well understood, that it is not yet entirely obvious how the components are created (5).

Bellow are few examples independent components and labels assigned to each of them. We list up to 20 most responsive words for each component to provide information for human judgment. These are examples from the 100 independent components experiment. For future research, perhaps an automatic way of determining different number of labels required to explain each independent component might be proposed using time series analysis, but for that, more research has to be provided to better understand the nature of independent components in order to justify such step.

First ten independent components can be seen in 1. As we can see, independent components cannot we regarded as synsets as known in WordNet, since they clearly contain words from multiple classes. We can perhaps call them collocation sets, colsets. But this term will have to be revised based on the subsequent research on the nature of independent components.

Table 4.1 shows an example how a particular word type is encoded. The independent components in this example are are sorted by the most important features. We can see how the encoding in Table 4.1 contrasts with Table 4.1, which shows ten least salient features for word type *yuyan* 語言.

4.2 Sense clustering

We have used maximum-linkage hierarchical algorithm from Pycluster package[4] to cluster word token contexts. The use of hierarchical clustering is motivated by the attempt to provide gradual sense analysis where subsenses could be identified within partial senses.

Our goal in this paper is to observe whether ICA based feature vectors can be efficiently used for word context encoding and subsequently for clustering. Clustering results were evaluated by native speaker with linguistics knowledge, who labelled all the sentences according to Chinese Wordnet and in this paper,

[4] http://bonsai.ims.u-tokyo.ac.jp/ mdehoon/software/cluster/software.htm

Label	IC	Responsive words (descending order)
TIME	0	時間 年 月 小時 天 段 半 經過 期間 週 分鐘 後 星期 久 持續 内 日 之後 工作 結束
TIME	1	三十 二十 五十 一百 四十 十 公尺 公里 歲 十五 以上 六十 超過 約 大約 左右 · 分鐘 十二 八十
FAMILY	2	媽媽 母親 孩子 爸爸 父親 女兒 父母 歲 兒子 家 小孩 弟弟 回家 妹妹 哥哥 帶 太太 照顧 回來 家人
COMPARATIVE	3	項 不同 電腦 好 什麼 系統 孩子 以上 後 路段 肯 設置 系所 當地 考 救 參與 最近 專線 事件
POPULATION	4	成長 去年 增加 今年 人數 減少 營收 達 預估 成長率 季 估 比例 期 衰退 同 高達 明年 人口 營業額
GAIN	5	得到 獲得 受到 受 肯定 給予 尊重 給 重視 關心 鼓勵 能 支持 表現 關懷 意見 太 都 獲 照顧
MULTIMEDIA	6	媒體 電視 新聞 廣告 報導 節目 雜誌 報紙 廣播 記者 電台 傳播 電視台 宣傳 製作 電子 大衆 電話 報 刊登
WAR	7	伊拉克 軍事 飛彈 部隊 攻擊 美國 中共 戰爭 蘇聯 武器 科威特 波斯灣 行動 聯合國 美 美軍 以色列 國防部 海珊 中東
WARNING	8	注意 應 不要 特別 避免 重要 應該 結果 提醒 最好 要 點 選擇 準備 安全 小心 健康 保持 飲食 呼籲
COMPETITION	9	選手 比賽 冠軍 運動 屆 中華 錦標賽 女子 亞運 參加 世界 協會 金牌 體育 男子 球員 國 我國 國際 教練
PRODUCTION	10	生產 技術 工業 製造 設備 工廠 產業 科技 產品 機械 材料 電子 廠 研發 化學 原料 知識 科學 加工 農業
ECONOMY	11	元 經費 費用 補助 預算 筆 錢 美元 負擔 金額 支出 收入 支付 新台幣 成本 貸款 每 資金 給 花費
RESEARCH	12	資料 調查 報告 結果 統計 顯示 分析 做 研究 份 進入 依據 指出 數據 預測 專家 地震 發現 評估 正確

Table 1: Independent components: 100 IC set, first 10 IC

Feature strength	Responsive words (descending order)
7.55055952072	用 字 聽 語言 首 英文 句 唱 音樂 詞 表達 歌 心 國語 獲得 寫 聲音 使用 詩 歌曲
6.93665552139	特色 具有 具 原住民 特殊 文化 獨特 語言 風格 色彩 豐富 背景 不同 特性 表現 很多 當地 傳統 歷史 最
6.20834875107	教學 英語 國小 國中 教育 學習 老師 課程 教師 小學 高中 學校 孩子 小朋友 學生 家長 教材 數學 教科書 英文
3.42819428444	她 得 我 他 快 孩子 玩 吃 深 態度 全 起來 父親 父母 跑 家庭 母親 共同 相當 一起
3.34706568718	品質 提高 高 提升 水準 成本 降低 效率 達到 低 提昇 改善 安全 整體 保障 服務 過 國民 享受 考量

Table 2: Partial example of encoded word 語言 (five most salient features)

Feature strength	Responsive words (descending order)
0.157807931304	申請 規定 昨天 不得 下午 取得 法院 任何 辦理 證明 行為 許 同意 違反 上午 是否 接受 機關 多 凌晨
0.157353967428	了解 不同 觀察 去 思考 看 分析 重新 調整 看看 深入 調查 重要 較 瞭解 面對 一下 探討 從 體會
0.152415782213	起 九月 三月 七月 六月 一日 五月 四月 二月 十二月 自 十月 民國 八月 至 十一月 底 一月 止 十五日
0.0953392237425	很 最 非常 相當 較 太 比較 更 十分 愈 比 越 極 得 那麼 這麼 一點 愈來愈 越來越 甚
0.0753756538033	選手 比賽 冠軍 運動 屆 中華 錦標賽 女子 亞運 參加 世界 協會 金牌 體育 男子 球員 國 我國 國際 教練

Table 3: Partial example of encoded word 語言 (five least salient features)

犯規 IC^{100}		
Cluster	Sense	Count
a	0	5
	1	28
b	0	3
	1	1

犯規 IC^{1000}		
Cluster	Sense	Count
a	0	1
	1	0
b	0	6
	1	30

Table 4: Results for word 犯規

措辭 IC^{100}		
Cluster	Sense	Count
a	0	9
	1	1
b	0	1
	1	8

措辭 IC^{1000}		
Cluster	Sense	Count
a	0	0
	1	1
b	0	9
	1	9

Table 5: Results for word 措辭

number of sense were also determined this way. Then we have assigned sense label to each cluster according to most prevalent sense in the cluster.

For example, word *fangui* 犯規 has two sense in Chinese Wordnet. We cut the tree produced by hierarchical clustering algorithm into two and our expectation is that word tokens manually labelled as sense 1 will be in one of the clusters and word tokens labelled as sense 2 will be in the other. Naturally some incorrect classifications can be expected as well and therefore we assign sense label according to the label most frequent in the particular cluster. In case we get both clusters labelled the same, the sense induction has failed.

In this experiment we have not pursued correct classification of all the words, therefore we leave the evaluation of those results out.

For reference we include tables with results of several words.

約談 IC^{100}		
Cluster	Sense	Count
a	0	22
	1	67
b	0	1
	1	12

約談 IC^{1000}		
Cluster	Sense	Count
a	0	1
	1	0
b	0	21
	1	80

Table 6: Results for word 措辭

山頭 IC^{100}		
Cluster	Sense	Count
a	0	39
	1	21
	2	11
b	0	7
	1	10
	2	0
c	0	2
	1	1
	2	0

山頭 IC^{1000}		
Cluster	Sense	Count
a	0	32
	1	39
	2	9
b	0	0
	1	3
	2	0
c	0	5
	1	2
	2	1

Table 7: Results for word 措辭

Word	IC100	IC1000
山頭	0	0
名牌	0	0
犯規	1	1
約談	0	1
措辭	1	1
堅硬	0	1
富有	0	0
屬下	0	1
天氣	0	0

Table 8: Overall results

5 Conclusion

The major advantage of our approach is that it uses global characteristics of words based on their co-occurrence with other words in the language, which are then applied to derive local encoding of word context. Thus we retrieve reliable characteristics of word's behaviour in the language and don't loose the word sense information, which allows us to analyze semantic characteristics of similar word forms.

Our current results are not very satisfying. I can be observed, however, from Table8 that increased number improves the sense induction considerably. We will pursue this track in our subsequent research. On the other hand, this result is not surprising. Considering the nature of independent components, which are rather symbolic features similar to synonymic sets, synsets, or rather collocation sets, collsets, it can be expected that much larger number of these components would be required to encode semantic information.

6 Future work

With manually semantically tagged word tokens we will try to automatically estimate the sufficient number of independent components that would improve precision of sense clustering.

Another approach we intend to try is to add feature vectors of all the context words and cluster the resulting vectors. This approach should emphasize more important features in given contexts.

We will also do more carefull preprocessing and also apply dimensionality reduction (typically done by PCA) before running ICA as has been done in some of the previous studies.

References

[1] E. Bingham. *Advances in Independent Component Analysis with Applications to Data Mining*. PhD thesis, Helsinki University of Technology, 2003.

[2] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250, New York, NY, USA, 2001. ACM Press.

[3] S. Bordag. Word sense induction: Triplet-based clustering and automatic

evaluation. In *11 th Conference of the European Chapter of the Association for Computational Linguistics: EACL 2006*, pages 137–144, 2006.

[4] T. Honkela and A. Hyvärinen. Linguistic feature extraction using independent component analysis. In *Proc. of IJCNN 2004*, 2004.

[5] T. Honkela, A. Hyvärinen, and J. Väyrynen. Emergence of linguistic features: Independent component analysis of context. In A. C. et al., editor, *Proceedings of NCPW9*, 2005.

[6] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*. Wiley, 2001.

[7] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural networks*, 13(4):411–430, 2001.

[8] S. Kaski. Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Proceedings of IJCNN'98, International Joint Conference on Neural Networks*, volume 1, pages 413–418. IEEE Service Center, Piscataway, NJ, 1998.

[9] T. Landauer and S. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 2001.

[10] D. B. Neill. Fully automatic word sense induction by semantic clustering. Master's thesis, Cambridge University, 2002.

[11] R. Rapp. A practical solution to the problem of automatic word sense induction. In *The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics*, pages 194–197, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[12] J. Väyrynen and T. Honkela. Comparison of independent component analysis and singular value decomposition in word context analysis. In *AKRR'05*, pages 135–140.