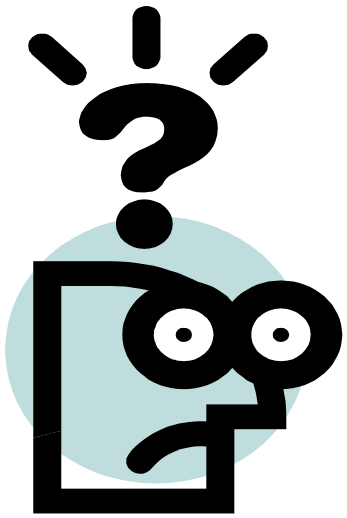
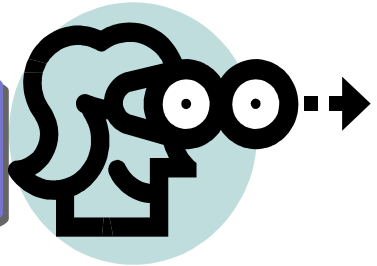
Support Vector Machine Based Orthographic Disambiguation

Eiji ARAMAKI, Takeshi IMAI, Kengo MIYO, Kazuhiko
THE UNIVERSITY OF TOKYO
Hospital



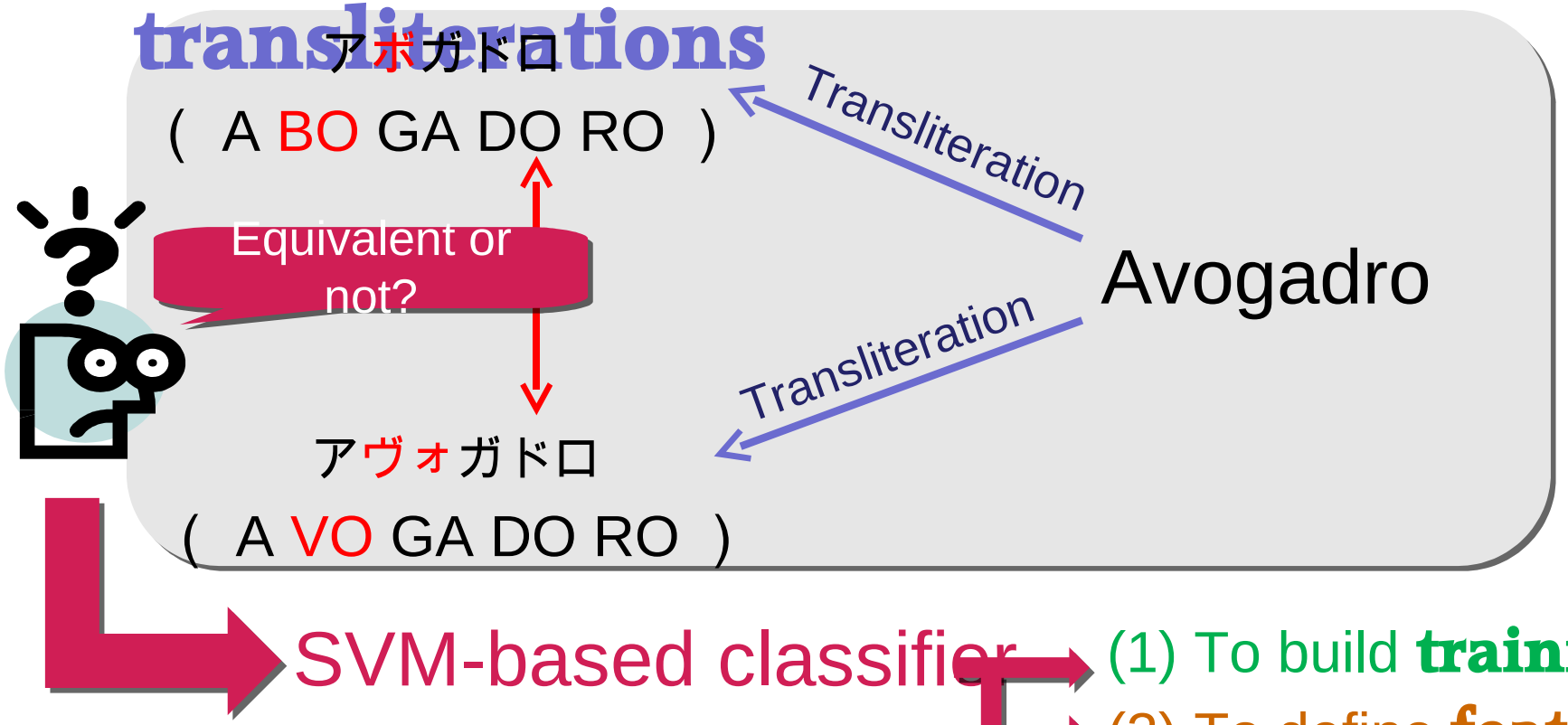
"center" and "centre" are equivalent?

We focus on Japanese, but the proposed method does not depend on languages



Background

- Japanese in particular contains orthographic variation, because of tons of **transliterations**



(1) Training-set

- **Positive** example: a term pair, which are spelled differently, but have the same meaning

in multiple translation dictionaries

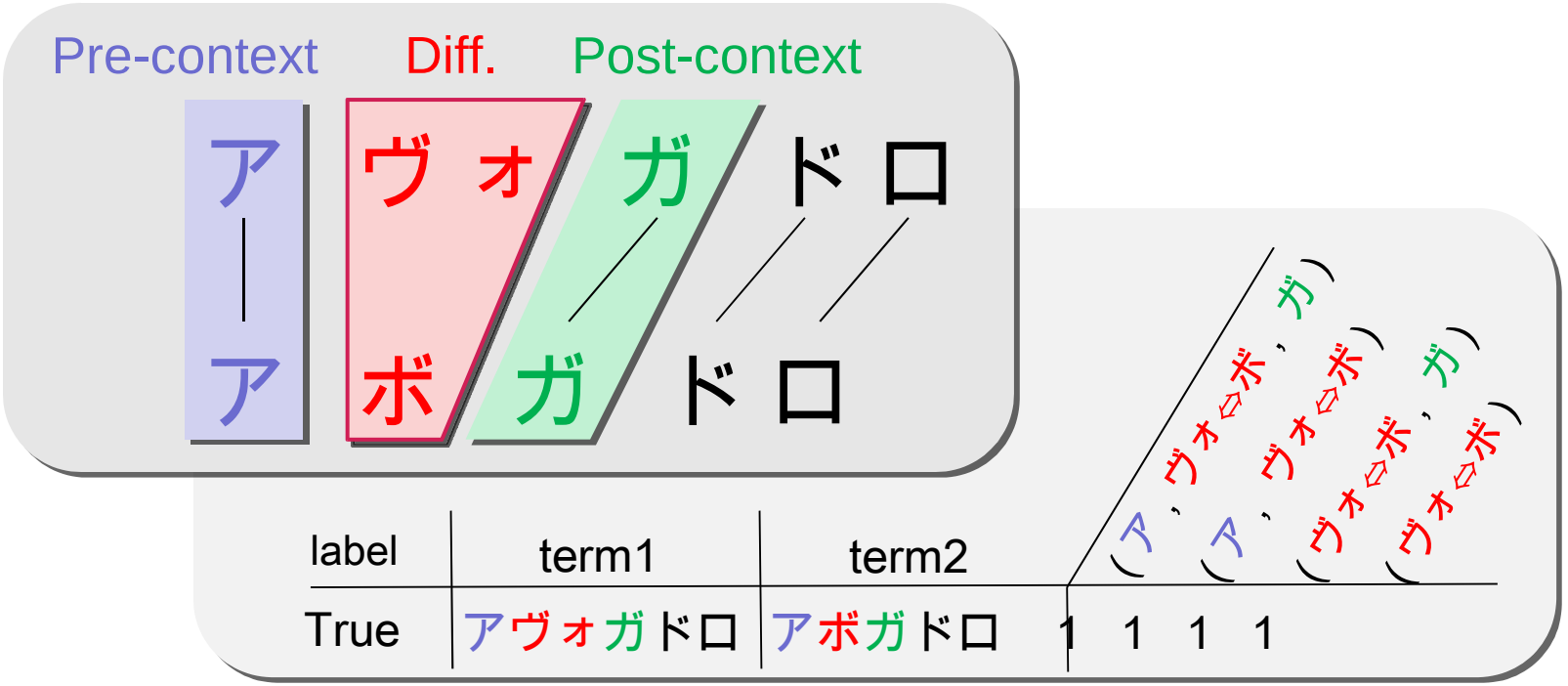
Same English Translation

- **Negative** example: a term pair, which are spelled differently and have different meanings.

Different English Translation

(2) Features for SVM

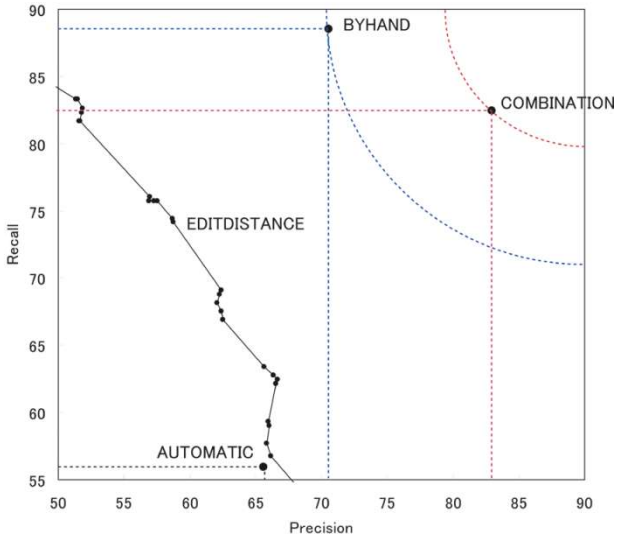
- different characters & its surrounding characters (window size=1; pre-context & post-context)



- Their combinations = **features**

Experiments

- Test-set: 883 Medical term pairs (312 positive)
- Methods:
  - (1) EDIT DISTANCE (*th*): IF SIM > *th* THEN +1
  - (2) BYHAND: SVM + 4,130 handmade training-set
  - (3) AUTOMATIC: SVM + 68,608 automatically built training-set
  - (4) COMBINATION: SVM + all training-set (BYHAND+AUTO
- Results:



$$Precision = \frac{\#\text{ of pairs found and correct}}{\text{total }\#\text{ of pairs found}},$$

$$Recall = \frac{\#\text{ of pairs found and correct}}{\text{total }\#\text{ of pairs correct}},$$

$$Accuracy = \frac{\#\text{ of pairs correct}}{\text{total }\#\text{ of pairs in test-set}}.$$

methods	Precision	Recall	Accuracy
EDIT-DISTANCE(0.91)	67.2%(164/244)	52.6% (164/312)	70.9% (626/883)
BYHAND	70.4%(276/392)	**88.4% (276/312)**	82.7% (731/883)
AUTOMATIC	65.7%(177/269)	56.7% (177/312)	74.2% (656/883)
COMBINATION	**82.9%(258/311)**	82.6% (258/312)	**87.8% (776/883)**

* The performance in EDIT-DISTANCE(0.91) showed the highest accuracy in various TH values.

Conclusion

- Discussion
  - Why AUTOMATIC < BYHAND
    - Because of Corpus specific styles (hepatitis-B or Hepatitis=B)
    - BYHAND corpus = test-set corpus ≠ AUTOMATIC corpus
- Conclusion
  - We proposed a discriminative orthographic disambiguation method.
  - We also proposed a method for collecting both positive & negative examples.
  - Experimental results yielded high levels of accuracy (87.8%), demonstrating the feasibility of the proposed approach.

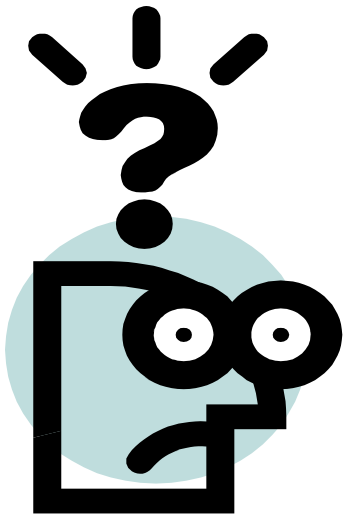
Unfortunately Bergsma [ACL2007] proposed similar methods

In the future, we will employ more features to boost the accuracy

P/N*	$Term_1$	$Term_2$
+1	ヨードピラセット (YO O DO PI RA SE TTO; iodopyracet)	ヨードピラセト (YO O DO PI RA SE TO; iodopyracet)
+1	マイクロメーター (MA I KU RO ME E TA A; micrometer)	マイクロメータ (MA I KU RO ME E TA; micrometer)
+1	アンプリファイア (A N PU RI FA I A; amplifier)	アンプリファイヤー (A N PU RI FA I YA A; amplifier)
+1	オシロスコープ (O SI RO SU KO O PU; oscilloscope)	オッシロスコープ (O SSI RO SU KO O PU; oscilloscope)
+1	動コンプライアンス (DO U KO N PU RA I A N SU; dynamic compliance)	動的コンプライアンス (DO U TE KI KO N PU RA I A N SU; dynamic compliance)
+1	浸透圧性ショック (SI N TO O A TU SE I SYO KKU; osmotic shock)	浸透圧ショック (SI N TO O A TU SYO KKU; osmotic shock)
-1	B 型肝炎 (BI I GA TA KA N E N; hepatitis B)	C 型肝炎 (SI I GA TA KA N E N; hepatitis C)
-1	トランス (TO RA N SU; trance)	トランジスタ (TO RA N JI SU TA; transistor)
-1	ビタミン P (BI TA MI N PI I; vitamin P)	ビタミン C (BI TA MI N SI I; vitamin C)
-1	カドミウム (KA DO MI U MU; cadmium)	カルシウム (KA RU SI U MU; calcium)
-1	アルコール (A RU KO O RU; alcohol)	グルコース (GU RU KO O SU; glucose)
-1	メラトニン (ME RA TO NI N; melatonin)	セロトニン (SE RA TO NI N; serotonine)

Support Vector Machine Based Orthographic Disambiguation

Eiji ARAMAKI, Takeshi IMAI, Kengo MIYO, Kazuhiko
THE UNIVERSITY OF TOKYO Hospital



"term1" and "term2" are equivalent?

We focus on Japanese, but the proposed method does not depend on languages