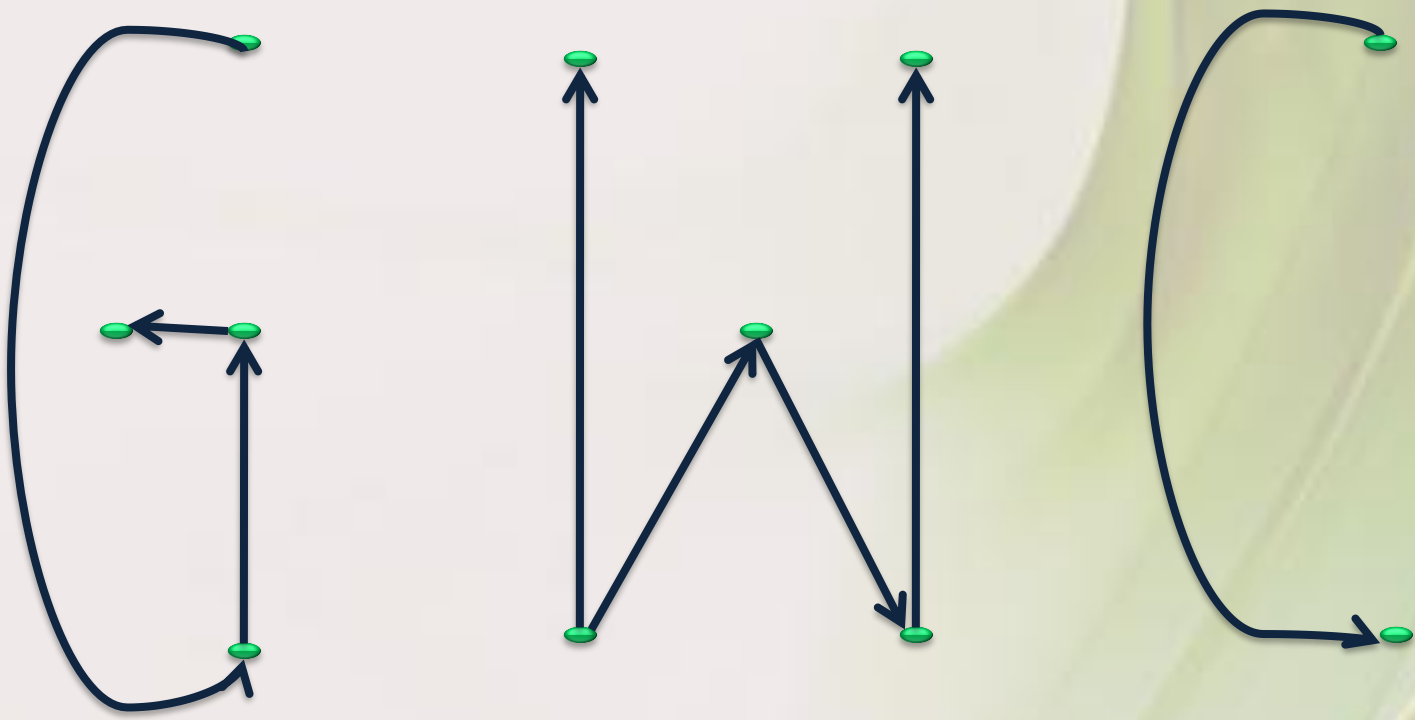
A Study of the Sense Annotation Process: Man v/s Machine

ARINDAM CHATTERJEE, SALIL JOSHI, PUSHPAK BHATTACHARYYA, DIPTESH KANOJIA AND AKHLESH MEENA

RoadMap



The Initial Step: IWSD Error Analysis

Iterative WSD (IWSD)

IWSD Scoring Function:

$$S^* = \arg\max_i \theta_i V_i + \sum_{j \in J} W_{ij} V_i V_j$$

STATISTICAL

CONTEXTUAL

- Which parameter contributes more?
- The Accuracy of IWSD has been constant in spite of several efforts.....why?

List of Experiments

To find the governing parameter for IWSD

- **Comparison between accuracies of IWSD and MFS**
- **Ablation Test on IWSD Parameters**
- **Suitable linear combination parameter (α)test:**

$$S^* = \alpha \arg\max_i \theta_i V_i + (1-\alpha) \sum_{j \in J} W_{ij} V_i V_j$$

Corpus Statistics (news)

	Polysemous words	Monosemous words	Wordnet Polysemy	Corpus Polysemy
Noun	72225	61682	3.03	1.82
Verb	26436	4372	4.47	3.00
Adj	15462	30122	2.68	2.03
Adv	12907	10658	2.52	2.11
Overall	127030	106834	3.13	2.02

Prime Parameter Test Results

MFS v/s IWSD

	Precision	Recall	F-Score
		52	79.04
		8	79.21

- IWSD is very close to MFS output
- This indicates predominance of the *P(S|W)* parameter

Ablation Test

Ablation Parameter	Precision	Recall	F-Score
θ	79.61%	78.62%	79.11%
$P(S\|W)$	[illegible]%	58.84%	59.21%
[illegible]	[illegible]58%		79.07%
[illegible]	[illegible]51%		79.01%
[illegible]	[illegible]62%		79.11%

- Clear indication that the *P(S|W)* statistic is the prime parameter for IWSD
- Knowledge based parameters have an accuracy of 60% as compared to 80% for *P(S|W)*

Alpha(α) Test

Alpha (α)	Precision	Recall	F-score
0	59.59%	58.84%	59.21
0.00001	79.48%	78.49%	78.98
0.0001	79.50%	51%	79
0.001	79.50%		79.01
			79.01
			79.11
			79.11
			79.11
			79.11
0.75	79.61%	78.62%	79.11
1	79.59%	78.60%	79.1

- ***P(S/W)*** **parameter has highest predominance for IWSD**
- **The predominance is so high that even for alpha = 0.00001, 80% accuracy is reached**

Introduction

general Belief

FOR HUMANS

Without Context

Inclusion of Context

Reduces Cognitive Load

With Context

In our research CONTEXT means the Neighboring words around the Target word

Observed fact

FOR MACHINES

State-of-the-Art WSD Algorithms are **Supervised**

Knowledge Based Algorithms have **Low Accuracies**

Sense Annotated Corpora

$P(Sense|Word)$

Motivation

Word Sense Disambiguation

Computationally identifying Senses of words in a CONTEXT

The river flows into the sea

Target word : sea

Context word : river

WSD is inspired by the human sense disambiguation technique

By definition WSD as an AI system, apes the human sense disambiguation technique – STRONG AI

Over the years, knowledge based WSD algorithms, which follow this technique, have reported low accuracies, which strongly indicates that machines fail to capture senses, in the human way

Supervised algorithms do not use human annotation techniques, yet deliver highest accuracies

This made us raise a question on the foundation of the WSD task – The relevance of its definition

WSD from the perspective of WEAK AI. We explored the dichotomy of man and machine annotation techniques

STRONG AI	WEAK AI
Machines should use context ***as humans do***, for WSD	Machines should use context ***in some way*** for WSD

Conversely, we also wanted to see if humans can annotate the way machines do *i.e.*, without context

In our research, machine annotation refers to the state-of-the-art WSD algorithms

Through our research we intend to answer the following questions fundamental to sense annotation:

- Can humans annotate without context as machines do?
- Do machines need context for sense disambiguation as humans do?

Our Claim

Humans and machines both need Context for annotation, but use context differently

Tagging without context is cognitively challenging for humans and highly erroneous

Humans cannot annotate the way machines do

Machines only need good sense statistics for annotation and do not need context the way humans do

Machines get the contextual evidence factored in the $P(S|W)$ measure, from Human Context Sensitive corpora tagged using context.

Machines conform to the principle of WEAK AI with respect to sense annotation

The Design of Experiments

Corpora and Annotation Scenarios

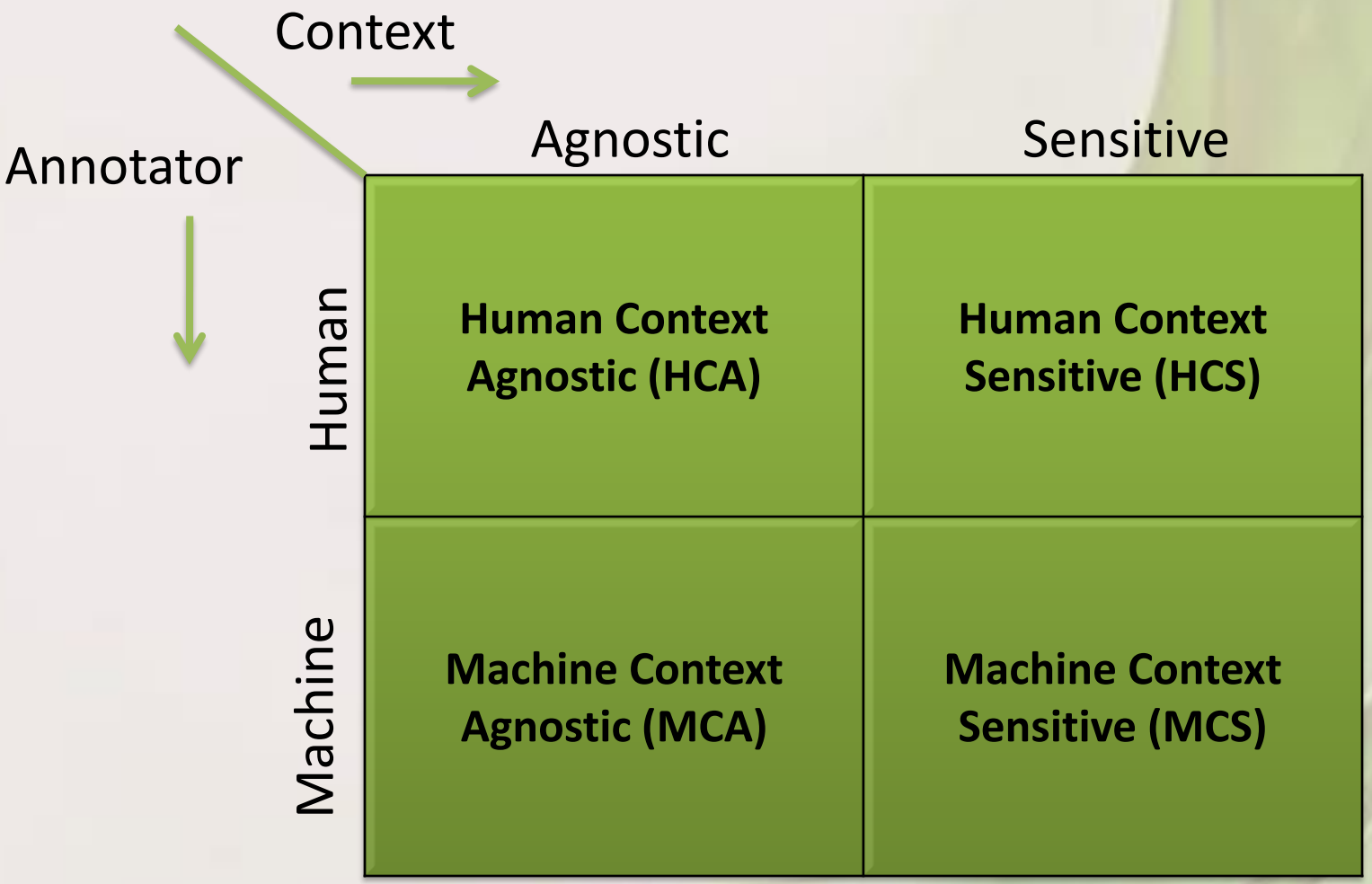
Context Sensitive Scenario

- **For Humans:** Specific domains (TOURISM and HEALTH) and generic domain (NEWS) were tagged *using context*
- **For Machines:** Trained and tested on context sensitive corpora

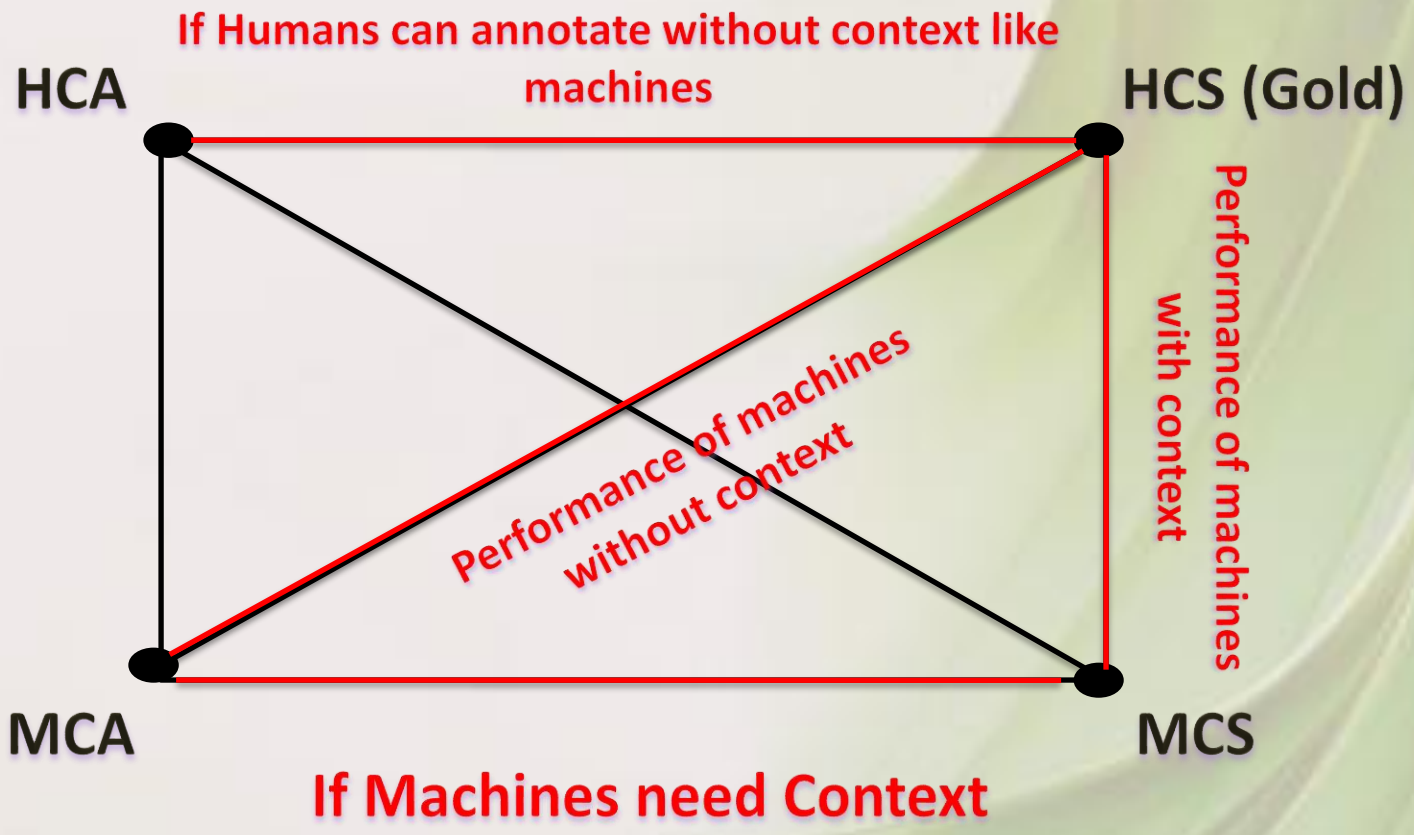
Context Agnostic Scenario

- **For Humans:** Specific domains (TOURISM and HEALTH) and generic domain (NEWS) were tagged without using context
  - The corpora used in this case consisted of a list of words, obtained from the corpora used in the context sensitive scenario
- **For Machines:** Trained and tested on context agnostic corpora

Annotation Genres



Why Annotation Genres?



Similarity Measures

Jaccard's Similarity Coefficient:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

where, A and B are annotators

Cohen's Kappa Coefficient:

$$\kappa(A, B) = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

where, *Pr(a)* is the relative observed agreement among annotators, and *Pr(e)* is the hypothetical probability of chance agreement

Human Annotation and Context

List of Experiments

To find the prime parameter for human annotation

- **Comparison between Human Context Sensitive and Context Agnostic data**

To find if humans can annotate accurately without context like machines

- **Comparison between Human Context Sensitive and Context Agnostic data**

To find the cognitive load for humans associated with tagging in both the scenarios

- **The time required for annotation was recorded**
- **The annotators' views after the tagging process were also recorded**

Experimental Settings

- Two annotators were assigned for the task
- Both annotators were native speakers of Hindi and skilled in the annotation task
- Annotated data obtained from a third annotator was used as Gold data for the comparison
- In context sensitive annotation, the annotation was done using context as is usual
- In the context agnostic scenario, the annotators knew the domain of the data before hand
- Laboratory conditions and facilities were same during context agnostic and context sensitive tagging

HCS v/s HCA – Jaccard's Similarity coefficient

- Humans need context for annotation
- Context is the prime parameter for human annotation
- Humans cannot annotate in the way machines do

					rall	
					ERB	OVERALL
HCS v/s HCA	TOURISM	0.65	0.48	0.63	0.69	0.61
	HEALTH	0.64	0.45	0.76	0.73	0.61
	NEWS	0.57	0.27	0.74	0.26	0.50

Insights

Humans need only contextual evidence for annotation

Annotation in context agnostic scenarios is cognitively challenging and erroneous

Humans cannot annotate without contextual evidence as machines do

Machine Annotation and Context

List of Experiments

To find if machine uses context for annotation

- **Comparison of machine outputs for context sensitive and context agnostic scenarios**
- **Comparison of machine output for context sensitive data with Gold data**
- **Comparison of machine output for context agnostic data with Gold data**

MCS v/s MCA – Kohen's Kappa statistic

- Match is low
- Machines trained on context agnostic data obtain poorer sense statistics than that trained on context sensitive data

Experiment	Domain	NOUN		ADV	VERB	OVERALL
MCA v/s MCS	TOURISM	0.34	0.13	0.05	0.31	0.27
	HEALTH	0.26	0.16	0.30	0.29	0.24
	NEWS	0.25	0.04	0.24	0.19	0.17

MCS v/s MCA – Jaccard's Similarity coefficient

Type of Experiment		POS and Overall				
		NOUN	ADJ	ADV	VERB	OVERALL
MCS v/s MCA	TOURISM	0.72	0.56	0.61	0.74	0.68
	HEALTH	0.69	0.56	0.79	0.69	0.67
	NEWS	0.66	0.40	0.75	0.53	0.62

HCS v/s MCS – Jaccard's Similarity coefficient

- Inter annotator agreement is 80%
- MCS performs at par with humans when trained on context sensitive data

Experiment		POS and Overall				
		NOUN	ADJ	ADV	VERB	OVERALL
HCS v/s MCS	TOURISM	0.80	0.71	0.82	0.78	0.78
	HEALTH	0.81	0.80	0.88	0.60	0.81
	NEWS	0.84	0.77	0.86	0.70	0.80

HCS v/s MCA – Jaccard's Similarity coefficient

- Machines require good sense statistics for high accuracy
- Sense statistics gathered from context agnostic corpora is poor
- Good sense statistics comes from context sensitive corpora annotated *using* context.

		nd Overall				
				ADV	VERB	OVERALL
HCS v/s MCA	TOURISM	0.65	0.48	.63	0.69	0.61
	HEALTH	0.64	0.45	0.76	0.73	0.61
	NEWS	0.57	0.27	0.74	0.26	0.50

Insights

P(S|W) is the prime parameter for machines

P(S|W) learnt from context sensitive data gives better accuracy than context agnostic data

Accurate P(S|W) is learnt from the corpus which is annotated using contextual evidence. Thus context sensitivity in machines is an adaption of Human Context Sensitive annotation.

Machine require contextual information. Unlike humans, machines use context through *P(S|W) parameter,* hence machines conform to the principle of WEAK AI

Man v/s Machine

Comparison Based on Ontological Categories

Tourism: Ontological Categories – Jaccard's coefficient

Ontological Category	TOURISM			
	Count	MCS v/s MCA	HCS v/s MCS	HCS v/s HCA
Verb of State	972	0.43	0.95	0.33
Action	863	0.25	0.83	0.21
Anatomical	798	0.35	0.89	0.34
Relational	721	0.33	0.75	0.18

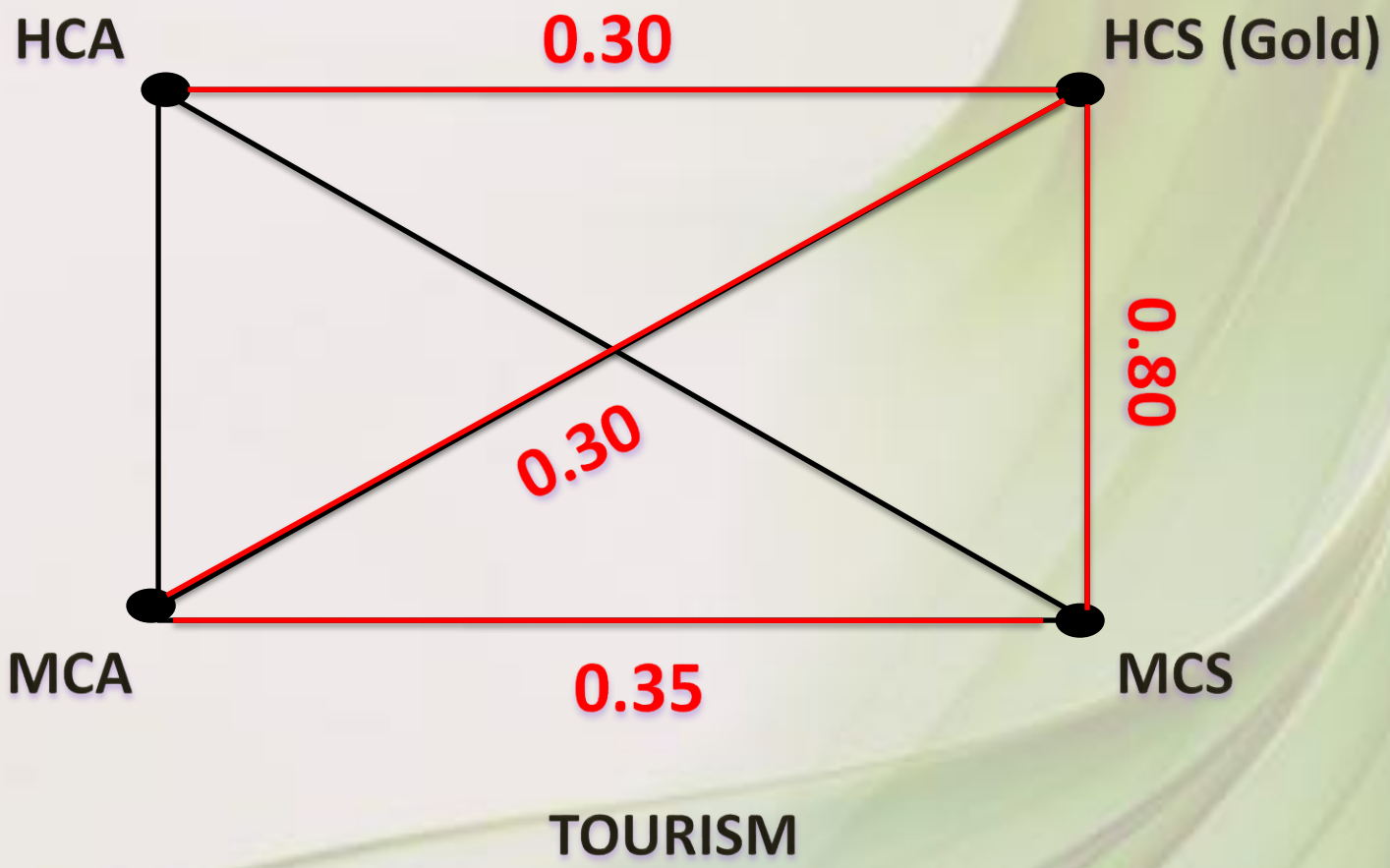
Health: Ontological Categories – Jaccard's coefficient

Ontological Category	HEALTH			
	Count	MCS v/s MCA	HCS v/s MCS	HCS v/s HCA
Bodily action	1198	0.06	0.95	0.89
Quantity	1188	0.01	0.90	0.16
Qualitative	1118	0.22	0.86	0.17
Numeral	1000	0.42	0.99	0.43

News: Ontological Categories – Jaccard's coefficient

Ontological Category	NEWS			
	Count	MCS v/s MCA	HCS v/s MCS	HCS v/s HCA
Physical Place	2209	0.67	0.92	0.73
Person	1829	0.47	0.90	0.70
Artifact	1796	0.27	0.85	0.61
Bodily action	1582	0.20	0.83	0.55

Insights (1/2)



Insights (2/2)

Humans need Context as the primary parameter for Annotation

Tagging without context is erroneous for humans. Humans cannot annotate like machines (*i.e.,* without context)

Machines need good sense statistics as the primary parameter for annotation

Machines require contextual information, but that is factored into the *P(S|W)* parameter, unlike humans' use of context

Conclusion

- Humans and machines both need Context for annotation, but use context differently ✓
- Tagging without context is cognitively challenging for humans and highly erroneous ✓
- Humans cannot annotate the way machines do ✓
- Machines only need good sense statistics for annotation ✓
- Machines get the contextual evidence factored in the P(S|W) measure, from Human Context Sensitive corpora tagged using context. ✓
- Word Sense Disambiguation is successful as a WEAK AI system. ✓

Future Work

For Humans: A deeper insight into the exact cognitive processes which are involved during the annotation process could further leverage the study between man and machine sense annotation processes.

Currently work is going on in this direction, by using an eye-tracking device to trace the exact use of context in human annotation

***For Machines*: Using better knowledge based parameters for IWSD could further enhance its accuracy.**

References

Lee, K. Yoong, Hwee T. Ng, and Tee K. Chia. 2004. *Supervised word sense disambiguation with support vector machines and multiple knowledge sources*. In Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, pages 137–140.

Eneko Agirre and German Rigau. 1996. Word sense disambiguation using conceptual density. In *In Proceedings of the 16th International Conference on Computational Linguistics (COLING).*

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. pages 92–100. Morgan Kaufmann Publishers.

Mitesh M. Khapra, Sapan Shah, Piyush Kedia, and Pushpak Bhattacharyya. 2009. Projecting parameters for multilingual word sense disambiguation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 459–467, Singapore,* August. Association for Computational Linguistics.

References

Mitesh Khapra, Saurabh Sohoney, Anup Kulkarni, and Pushpak Bhattacharyya. 2010. Value for money: Balancing annotation effort, lexicon building and accuracy for multilingual wsd. In *Proceedings of the 23rd International Conference on Computational Linguistics.*

Yoong Keok Lee, Hwee Tou Ng, and Tee Kiah Chia. 2004. Supervised word sense disambiguation with support vector machines and multiple knowledge sources. In *Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, pages 137–140.*

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *In Proceedings of the 5th annual international conference on Systems documentation.*

Hang Li and Cong Li. 2004. Word translation disambiguation using bilingual bootstrapping. *Comput. Linguist.,* 30:1–22,March.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, page 279, Morristown, NJ, USA.* Association for Computational Linguistics.

References

Rada Mihalcea. 2005. Large vocabulary unsupervised word sense disambiguation with graph-based algorithms for sequence data labeling. In *In Proceedings of the Joint Human Language Technology and Empirical Methods in Natural Language Processing Conference (HLT/EMNLP), pages 411–418.*

Rajat Mohanty, Pushpak Bhattacharyya, Prabhakar Pande, Shraddha Kalele, Mitesh Khapra, and Aditya Sharma. 2008. Synset based multilingual dictionary: Insights, applications and challenges. In *GlobalWordnet Conference.*

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word senses: An exemplar-based approach. In *In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL), pages 40–47.* D. Walker and R. Amsler. 1986. The use of machine readable dictionaries in sublanguage analysis. In *In Analyzing Language in Restricted Domains, Grishman and Kittredge (eds), LEA Press, pages 69–83.* David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics, pages 189–196, Morristown,* NJ, USA. Association for Computational Linguistics.

Thank You

? ? ?