A Generalized Framework for Revealing Analogous Themes across Related Topics

Zvika Marx
CS and AI Laboratory
MIT
Cambridge, MA 02139, US
zvim@csail.mit.edu

Ido Dagan
Computer Science Department
Bar-Ilan University
Ramat-Gan 52900, Israel
dagan@cs.biu.ac.il

Eli Shamir
School of Computer Science
The Hebrew University
Jerusalem 91904, Israel
shamir@cs.huji.ac.il

Abstract

This work addresses the task of identifying thematic correspondences across sub-corpora focused on different topics. We introduce an unsupervised algorithmic framework based on distributional data clustering, which generalizes previous initial works on this task. The empirical results reveal interesting commonalities of different religions. We evaluate the results through measuring the overlap of our clusters with clusters compiled manually by experts. The tested variants of our framework are shown to outperform alternative methods applicable to the task.

1 Introduction

The ability to identify analogies and correspondences is one of the fascinating aspects of intelligence. Research in cognitive science has acknowledged the significance of this ability of human thinking, particularly in learning across different situations or domains where the common base to learning is not straightforward. Several previous computational models of analogy making (e.g. Falkenhainer et al., 1989) suggested symbolic computational mechanisms for constructing detailed mappings that connect corresponding ingredients across analogized systems.

This work explores the identification of thematic correspondences in texts through an extension of the well known data clustering problem. Previous works aimed at identifying – through clusters of words – concepts, sub-topics or themes that are prominent within a corpus of texts (e.g., Pereira et al., 1993; Li, 2002; Lin and Pantel, 2002). The current work deals with extending this line of research to identify corresponding themes across a corpus pre-divided to several sub-corpora, which are focused on different, yet related, topics.

This research task has been defined quite recently (Dagan et al., 2002), and has not been explored extensively yet. One could think, however, of many potential applications for drawing correspondences across textual resources: comparison of related firms or products, identifying equivalencies in news published in different countries, and so on. The experimental part of our work deals with revealing correspondences between different religions: Buddhism, Christianity, Hinduism, Islam and Judaism. Given a pre-partition of the corpus to sub-corpora, one for each religion, our method exposes common aspects for all religions, such as *sacred writings*, *festival*s and *suffering*.

The mechanism we employ directs corresponding key terms in the different sub-corpora, such as names of festivals of different religions, to be included in the same cluster. Term clustering methods in general, and in this work in particular, rely on word co-occurrence statistics: terms sharing similar words co-occurrence statistics are clustered together. Different topics, however, are characterized by distinctive terminology and typical word co-locations. Therefore, given a pre-divided corpus, similar co-occurrence patterns would typically be extracted from the same topical sub-corpus. When the terminology and typical phrases employed by each topic differ greatly (even if the top-

ics are essentially related, e.g. different religions), the tendency to form topic-specific clusters intensifies regardless of factors that otherwise could have impact this tendency, such as the co-occurrence window size. Consequently, corresponding key terms of different topics may not be assigned by a standard method to the same cluster, in contrast to our goal. The method described in this paper aims precisely at this problem: it is designed to neutralize salient co-occurrence patterns within each topical sub-corpus and to promote less salient patterns that are shared across the sub-corpora.

In an earlier line of research we have formulated the above problem and addressed it within a probabilistic vector-based setting, presenting two related heuristic algorithms (Dagan et al., 2002; Marx et al., 2004). Here, we devise a general principled distributional clustering paradigm for this problem, termed *cross-partition clustering*, and show that the earlier algorithms are special cases of the new framework.

This paper proceeds as follows: Section 2 describes in more detail the cross-partition clustering problem. Section 3 reviews distributional data clustering methods, which form the basis to our algorithmic framework described in Section 4. Section 5 presents experimental results that reveal interesting themes common to different religions and demonstrates, through an evaluation based on human expert data, that the different variants of our framework outperform alternative methods.

2 The cross-partition clustering problem

The *cross-partition* clustering problem is an extension of the standard (single-set) data clustering problem. In the cross-partition setting, the dataset is pre-partitioned into several distinct *subsets* of elements to be clustered. For example, in our experiments each of these subsets consisted of topical key terms to be clustered. Each such subset was extracted automatically from a sub-corpus corresponding to a different religion (see Section 5).

As in the standard clustering problem, our goal is to cluster the data such that each term cluster would capture a particular theme in the data. However, the generated clusters are expected to identify themes that cut *across* all the given subsets. For example, one cluster consists of names of festivals of different religions, such as *Easter*, *Christmas*, *Sunday* (Christianity) *Ramadan*, *Friday*, *Id-al-fitr* (Islam) and *Sukoth*, *Shavuot*, *Passover* (Judaism; see Figure 4 for more examples).

3 Distributional clustering

Our algorithmic framework elaborates on Pereira et al.'s (1993) *distributional clustering* method. Distributional clustering probabilistically clusters data elements according to the distribution of a given set of features associated with the data. Each data element x is represented as a probability distribution $p(y|x)$ over all features y. In our data $p(y|x)$ is the empirical co-occurrence frequency of a feature word y with a key term x, normalized over all feature word co-occurrences with x.

The distributional clustering algorithmic scheme (Figure 1) is a probabilistic (soft) version of the well-known *K-means* algorithm. It iteratively alternates between:

(1) ***Calculating assignments to clusters***: calculate an assignment probability $p(c|x)$ for each data elements x into each one of the clusters c. This soft assignment is proportional to an information theoretic distance (KL divergence) between the element's $p(y|x)$ representation, and the centroid of c, represented by a distribution $p(y|c)$. The marginal cluster probability $p(c)$ may optionally be set as a prior in this calculation, as in Tishby et al. (1999; in Figure 1 we mark it with dotted underline, to denote it is optional).

Set $t = 0$, and repeatedly iterate the two update-steps below, till convergence (at time step $t = 0$, initialize $p_t(c|x)$ randomly or arbitrarily and skip step 1):

(1) $$p_t(c|x) = \underline{p_{t-1}(c)} \frac{e^{-\beta KL[p(y|x)\| p_{t-1}(y|c)]}}{z_t(x,\beta)}$$

where $$z_t(x,\beta) = \sum_{c'} \underline{p_{t-1}(c')} e^{-\beta KL[p(y|x)\| p_{t-1}(y|c')]}$$

(2) $$p_t(y|c) = \frac{1}{p_t(c)} \sum_x p(x) p_t(c|x) p(y|x)$$

where $$p_t(c) = \sum_x p(x) p_t(c|x)$$

(3) $$t = t + 1$$

Figure 1: A general formulation of the iterative distributional clustering algorithm (with a fixed β value and a fixed number of clusters). The underlined $\underline{p_{t-1}(c)}$ term is optional.

(2) ***Calculating cluster centroids***: calculate a probability distribution $p(y|c)$ over all features y given each cluster c, based on the feature distribution of cluster elements, weighed by the $p(c|x)$ assignment probability calculated in step (1) above. This step imposes the independence of the clusters c of the features y given the data x (similarly to the *naïve Bayes* supervised framework).

Subsequent works (Tishby et al., 1999; Gedeon et al., 2003) have studied and motivated further the earlier distributional clustering method. Particularly, it can be shown that the algorithm of Figure 1 locally minimizes the following *cost function*:

$$F^{dist\text{-}clust} \equiv \underline{H(C)} - H(C|X) + \beta H(Y|C), \qquad (1)$$

where H denotes entropy[1] and X, Y and C are formal variables whose values range over all data elements, features and clusters, respectively.

Tishby et al.'s (1999) *information bottleneck* method (IB) includes the marginal cluster entropy $\underline{H(C)}$ in the cost term[2] (it is marked with dotted underline to denote its inclusion is optional, so that Eq. (1) encapsulates two different cost terms). The addition of $\underline{H(C)}$ corresponds to including the optional prior term $\underline{p_{t-1}(c)}$ in step (1) of the algorithm.

The parameter β that appears in the cost term and in step (1) of the algorithm can have any positive real value. It counterbalances the relative impact of the considerations of maximizing feature information conveyed by the partition to clusters, i.e. minimizing $H(Y|C)$, versus applying the *maximum entropy principle* to the cluster assignment probabilities (see Gedeon et al., 2004), i.e., maximizing $H(C|X)$. The higher β is, the more "determined" the algorithm becomes in assigning each element into the most appropriate cluster. In subsequent runs of the algorithm β can be increased, yielding more separable clusters (clusters with noticeably different centroids) upon convergence. The runs can repeat until, for some β, the desired number of separate clusters is obtained.

4 The cross-partition clustering method

In the cross-partition framework, the pre-partition of the data to subsets is captured through an additional formal variable W, whose values range over the subsets. In our data, each religion corresponds to a different W value, w. Each religion-related key term x is associated with one religion w, with $p(w|x) = 1$ (and $p(w'|x) = 0$ for any $w' \neq w$). Formally, our framework allows probabilistic pre-partition, i.e., $p(w|x)$ values between 0 and 1 but this option was not examined empirically.

The Cross-Partition (CP) clustering method (Figure 2) is an extended version of the probabilistic K-means scheme, introducing additional steps in the iterative loop that incorporate the added pre-partition variable W:

(1) ***Calculating assignments to clusters***, i.e. probabilistic $p(c|x)$ values, is based on current values of cluster *centroids*, as in distributional clustering.

(2) ***Calculating subset-projected cluster centroids***. Given the current element assignments, centroids are computed separately for each combination of

Set $t = 0$ and repeatedly iterate the following update steps sequence, till convergence (in the first iteration, when $t = 0$ randomly or arbitrarily initialize $p_t(c|x)$ and skip step CP1):

(1) $$p_t(c|x) = \underline{p_{t-1}(c)} \frac{e^{-\beta KL[p(y|x)\|p^*_{t-1}(y|c)]}}{z_t(x,\beta)}$$
where $$z_t(x,\beta) = \sum_{c'} \underline{p_{t-1}(c')} e^{-\beta KL[p(y|x)\|p^*_{t-1}(y|c')]}$$

(2) $$p_t(y|c,w) = \frac{1}{p_t(c,w)} \sum_x p(x) p_t(c|x) p(y|x) p(w|x)$$
where $$p_t(c,w) = \sum_x p(x)\, p_t(c|x)\, p(w|x)$$

(3) $$p^*_t(c|y) = \underline{p^*_{t-1}(c)} \frac{\prod_w p_t(y|c,w)^{\eta p(w)}}{z^*_t(y,\eta)}$$
where $$z^*_t(y,\eta) = \sum_{c'} \underline{p_{t-1}(c')} \prod_w p_t(y|c',w)^{\eta p(w)}$$

(4) $$p^*_t(y|c) = \frac{1}{p^*_t(c)} \sum_y p(y)\, p^*_t(c|y)$$
where $$p^*_t(c) = \sum_y p(y)\, p^*_t(c|y)$$

(5) $$t = t + 1$$

Figure 2: The cross partition clustering iterative algorithm (with fixed β and η values and a fixed number of clusters). The terms marked by dotted underline, $\underline{p_{t-1}(c)}$ and $\underline{p^*_t(c)}$, are optional.

[1] The entropy of a random variable A is $H(A) = \sum_a p(a) \log p(a)$, where a ranges over all values of A; the entropy of A conditioned on another variable B is $H(A|B) = \sum_{a,b} p(a,b) \log p(a|b)$, with a and b range over all values of A and B.

[2] The IB cost function was originally formulated as $F^{IB} \equiv I(C;X) - \beta I(C;Y)$. This formulation is equivalent to ours, as $I(C;X) = H(C) - H(C|X)$ and $I(C;Y) = H(Y) - H(Y|C)$, while $H(Y)$ is a constant term depending only on the data.

a cluster c projected on a pre-given subset w. Each such subset-projected centroid is given by a probability distribution $p(y|c,w)$ over the features y, for each c and w separately (instead of $p(y|c)$).

(3) ***Re-evaluating cluster-feature association***. Based on the subset projected centroids, the associations between features and clusters are re-evaluated: features that are commonly prominent across all subsets are promoted relatively to features with varying prominence. A weighted geometric mean scheme achieves this effect: the value of $\prod_w p(y|c,w)^{\eta p(w)}$ is larger as the different $p(y|c,w)$ values are distributed more uniformly over the different w's, for any given c and y. η is a positive valued free parameter, which controls the impact of uniformity versus variability of the averaged values. The re-evaluated associations resulting from this stage are probability distributions over the clusters denoted $p^*(c|y)$. We add an asterisk to distinguish this conditioned probability distribution from other $p(c|y)$ values that can be calculated directly from the output of the previous steps.

(4) ***Calculating cross-partition "global" centroids***: based on the probability distributions $p^*(c|y)$, we calculate a probability distribution $p^*(y|c)$ for every cluster c through a straightforward application of Bayes rule, obtaining the cross partition cluster centroids.

The novelty of the CP algorithm lies in step (3): rather than deriving cluster centroids directly, as in the standard k-means scheme, cluster-feature associations are biased by their prominence *across* the cluster projections over the different subsets. This way, only features that are prominent in the cluster across most subsets end up prominent in the eventual cluster centroid (computed in step 4). By incorporating for every c–y pair a product over all w's, independence of the feature-cluster associations from specific w values is ensured. This conforms to our target of capturing themes that cut across the pre-given partition and are not correlated with specific subsets.

Employing a separate update step in order to accomplish the above direction implies deviation from the familiar cost-based scheme. Indeed, the CP method is not directed by a single cost function that globally quantifies the cross partition clustering task on the whole. Rather, there are four different "local" cost-terms, each articulating a different aspect of the task. As shown in the appendix, each of the update steps (1)–(4) reduces one of these four cost terms, under the assumption that values not modified by that step are held constant. This assumption of course does not hold as values that are not modified by a given step are modified by another. Hence, downward convergence (of any of the cost terms) is not guaranteed.

However, empirical experimentation shows that the dynamics of the CP algorithm tend to stabilize on an equilibrial steady state, where the four different distributions produced by the algorithm balance each other, as illustrated in Figure 3. In fact, convergence occurred in all our text-based experiments (as well as in experiments with synthetic data; Marx et al., 2004).

Manipulating the value of the β parameter works in practice for the CP method as it works for distributional clustering: increasing β along subsequent runs enables the formation of configurations of growing numbers of clusters. The CP framework introduces an additional parameter, η. Intuitively, step (3). As said, the geometric mean scheme promotes those c–y associations for which the $p(y|c,w)$ values are distributed evenly across the w's (for any fixed c and y). A low η would imply a relatively low penalty to those c–y combinations that are not distributed evenly across the w's, but it

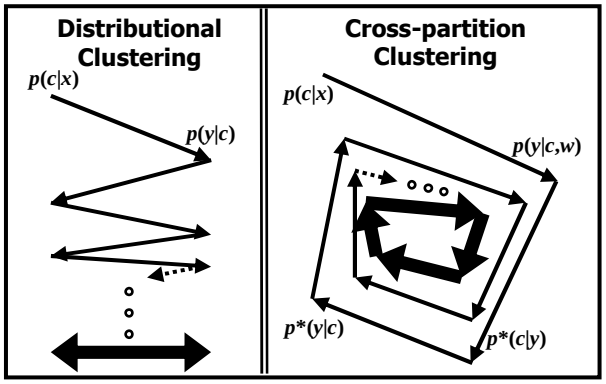

Figure 3: A schematic illustration of the dynamics of the CP framework versus that of distributional clustering. In distributional clustering convergence is onto a configuration where the two systems of distributions complementarily balance one another, bringing a cost term to a locally minimal value. In CP, stable configurations maintain balanced inter-dependencies (*equilibrium*) of four systems of probability distributions.

entails also loss of more information compared to high η. We experimented with η values that are fixed during a whole sequence of runs, while only β is gradually incremented (see Section 5).

Likewise the optional incorporation of priors in the distributional clustering scheme (Figure 1), the CP framework detailed in Figure 2 encapsulates four different algorithmic variants: the prior terms (marked in Figure 2 with dotted underline) can be optionally added in steps (1) and/or (3) of the algorithm. As in the distributional clustering case, the inclusion of these terms corresponds to inclusion of cluster entropy in the corresponding cost terms (see Appendix). It is interesting to note that we introduced previously, on intuitive accounts, some of these variants separately. Here we term the three variations involving priors CP_I (prior added in step (1) only, which is the same as the method described in Dagan et al., 2002), CP_{II} (prior added in step (3) only) and CP_{III} (prior added in both steps; as the method in Marx et al., 2004). The version with no priors is denoted CP. Our formulation reveals that these are all special cases of the general CP framework described above.

5 Experimental Results

The data elements that we used for our experiments – religion related key terms – were automatically extracted from a pre-divided corpus addressing five religions: Buddhism, Christianity, Hinduism, Islam and Judaism. The clustered key-term set was pre-partitioned, correspondingly, to five disjoint subsets, one per religion w.[3] In our experimental setting, the key term subsets for the different religions were considered disjoint, i.e., occurrences of the same word in different subsets were considered distinct elements. The set of features y consisted of words that co-occurred with key terms within ±5 word window, truncated by sentence boundaries. About 7000 features, each occurring in all five sub-corpora, were selected.

We survey below some results, which were produced by the plain (unprioired) CP algorithm with $\eta = 0.48$ applied to all five religions together. First, we describe our findings qualitatively and afterwards we provide quantitative evaluation.

[3] We use the dataset of Marx et al. (2004) – five sub-corpora, of roughly one million words each, consisting of introductory web pages, electronic journal papers and encyclopedic entries about the five religions; about 200 key terms were extracted from each sub-corpus to form the clustered subsets.

5.1 Cross-religion Themes

We have found that even the coarsest partition of the data to two clusters was informative and illuminating. It revealed two major aspects that seem to be equally fundamental in the religion domain. We termed them the "spiritual aspect" and "establishment aspect" of Religion. The "spiritual" cluster incorporated terms related with theology, underlying concepts and personal religious experience. Many of the terms assigned to this cluster with highest probability, such as *heaven*, *hell*, *soul*, *god* and *existence*, were in common use of several religions, but it included also religion-specific words such as *atman*, *liberation* and *rebirth* (key concepts of Hinduism). The "establishment" cluster contained names of schools, sects, clergical positions and other terms connected to religious institutions, geo-political entities and so on. Terms assigned to this cluster with high probability were mainly religion specific: *protestant*, *vatican*, *university*, *council* in Christianity; *conservative*, *reconstructionism*, *sephardim*, *ashkenazim* in Judaism and so on (few terms though were common to several religions, for instance *east* and *west*). This two-theme partition was obtained persistently (also when the CP method was applied to pairs of religions rather than to all five). Hence, these aspects appear to be the two universal constituents of religion-related texts in general, to the level the data reflect faithfully this domain.

Clusters of finer granularity still seem to capture fundamental, though more focused, themes. For example, the partition into seven clusters revealed the following topics (our titles): "schools", "divinity", "religious experience", "writings", "festivals and rite", "material existence, sin, and suffering" and "family and education". Figure 4 details the members of highest $p(c|x)$ values within each religion in each of the seven clusters.

The relation between the seven clusters to the coarser two-cluster configuration can be described in soft-hierarchy terms: the "schools" cluster and, to some lesser extent "festivals" and "family", are related with the "establishment aspect" reflected in the partition to two, while "divinity", "religious experience" and "suffering" are clearly associated with the "spiritual aspect". The remaining topic, "writings", is equally associated with both. The probabilistic framework enabled the CP method to

CLUSTER 1 "Schools"
Buddhism: america asia japan west east korea india china tibet
Christianity: orthodox protestant catholic west orthodoxy organization rome council america
Hinduism: west christian religious civilization buddhism aryan social founder shaiva
Islam: africa asia west east sunni shiah christian country civilization philosophy
Judaism: reform conservative reconstructionism zionism orthodox america europe sephardim ashkenazim

CLUSTER 2 "Divinity"
Buddhism: god brahma
Christianity: holy-spirit jesus-christ god father savior jesus baptize salvation reign
Hinduism: god brahma
Islam: god allah peace messenger jesus worship believing tawhid command
Judaism: god hashem bless commandment abraham

CLUSTER 3 "Religious Experience"
Buddhism: phenomenon perception consciousness human concentration mindfulness physical liberation
Christianity: moral human humanity spiritual relationship experience expression incarnation divinity
Hinduism: consciousness atman human existence liberation jnana purity sense moksha
Islam: spiritual human physical moral consciousness humanity exist justice life
Judaism: spiritual human existence physical expression humanity experience moral connect

CLUSTER 4 "Writings"
Buddhism: pali-canon sanskrit sutra pitaka english translate chapter abhidhamma book
Christianity: chapter hebrew translate greek new-testament book text old-testament luke
Hinduism: rigveda gita sanskrit upanishad sutra smriti brahma-sutra scripture mahabharata
Islam: chapter surah bible write translate hadith book language scripture
Judaism: tanakh scripture mishnah book oral talmud bible write letter

CLUSTER 5 "Festivals and Rite"
Buddhism: full-moon celebration stupa ceremony sakya abbot ajahn robe retreat
Christianity: easter tabernacle christmas sunday sabbath jerusalem pentecost city season
Hinduism: puja ganesh festival ceremony durga rama pilgrimage rite temple
Islam: kaabah id ramadan friday id-al-fitr haj mecah mosque salah
Judaism: sukoth festival shavuot temple passover jerusalem rosh-hashanah temple-mount rosh-hodesh

CLUSTER 6 "Sin, Suffering, Material Existence"
Buddhism: lamentation water grief kill eat hell animal death heaven
Christianity: fire punishment eat water animal lost hell perish lamb
Hinduism: animal heaven earth death water kill demon birth sun
Islam: water animal hell punishment paradise food pain sin earth
Judaism: animal water eat kosher sin heaven death food forbid

CLUSTER 7 "Family and Education"
Buddhism: child friend son people family question learn hear teacher
Christianity: friend family mother boy question woman problem learn child
Hinduism: child question son mother family learn people teacher teach
Islam: sister husband wife child family marriage mother woman brother
Judaism: child marriage wife mother father women question family people

Figure 4: A sample from a seven-cluster CP configuration of the religion data, including the first members – up to nine – of highest $p(c|x)$ within each religion in each cluster. Cluster titles were assigned by the authors for reference.

cope with these composite relationships between the coarse partition and the finer one.

It is interesting to have a notion of those features y with high $p^*(c|y)$, within each cluster c. We exemplify those typical features, for each one of the seven clusters, through four of the highest $p^*(c|y)$ features (excluding those terms that function as both features and clustered terms):

- "schools" cluster:
  central, *dominant*, *mainstream*, *affiliate*;
- "divinity" cluster:
  omnipotent, *almighty*, *mercy*, *infinite*;
- "religious experience" cluster:
  intrinsic, *mental*, *realm*, *mature*;
- "writings" cluster:
  commentary, *manuscript*, *dictionary*, *grammar*;
- "festivals and rite" cluster:
  annual, *funeral*, *rebuild*, *feast*;
- "material existence, sin, and suffering" cluster:
  vegetable, *insect*, *penalty*, *quench*;
- "community and family" cluster:
  parent, *nursing*, *spouse*, *elderly*.

We demonstratively focus on the two-cluster and seven-cluster, as these numbers are small enough to allow review of all clusters. Configurations of more clusters revealed additional subtopics, such as *education*, *prayer* and so on.

There are some prominent points of correspondence between our findings to Ninian Smart's comparative religion classics *Dimensions of the Sacred* (1996). For instance, Smart's *ritual* dimension corresponds to our "festivals and rite" cluster and his *experiential and emotional* dimension corresponds to our "religious experience" cluster.

5.2 Evaluation with Expert Data

We evaluated the performance of our method against cross-religion key term clusters constructed manually by a team of three experts of comparative religion studies. Each manually produced clustering configuration referred to two of the five religions rather than to all five jointly, as in our qualitative review. We examined eight of the ten religion pairs that can be chosen from the total of

five. Each religion pair was addressed independently by two different experts using the same set of key terms (so the total number of contributed configurations was 16). Thus, we could also asses the level of agreement between experts.

As an overlap measure we employed the Jaccard coefficient, which is the ratio $n_{11}/(n_{11}+n_{10}+n_{01})$, where:

n_{11} is the number of term pairs assigned to the same cluster by both our method and the expert;

n_{10} is the number of term pairs co-assigned by our method but not by the expert;

n_{01} is the number of term pairs co-assigned by the expert but not by our method.

As the Jaccard score relies on counts of individual term pairs, no assumption with regard to the suitable number of clusters is required. Hence, for each religion pair we produced with our method configurations of two to 16 clusters and calculated for each Jaccard scores based on the overlap with the relevant expert configurations. The scores obtained were averaged over the 15 configurations. The means, over all 16 experimental cases, of those average values are displayed in Table 1.

We tested all four CP method variants, with different fixed values of the η parameter. In addition, we evaluated results obtained by the priored version of distributional clustering (the IB method, Tishby et al., 1999; see Figure 1). Marx et al. (2004) mentioned Information Bottleneck with Side Information (IB-SI, Chechik & Tishby, 2003) as a method capable – unlike standard distributional clustering – of capturing information regarding pre-partition to subsets, which makes this method a seemingly sensible alternative to the CP method. Therefore, we tested the IB-SI method as well, following the adaptation scheme to the CP setting described by Marx et al, with a fixed value of its parameter, $\gamma = 0.07$ (with higher values convergence did not take place in all experiments). As Table 1 shows, the different CP variants performed better than the alternatives. The CP_{III} varinat, with both prior types, was less robust to changes in η value and seemed to be more sensitive to noise.

The experimental part of this work demonstrates that the task of drawing thematic correspondences is challenging. In the particular domain that we have examined the level of agreement between experts seems to make it evident that the task is inherently subjective and just partly consensual. It is remarkable therefore that most variations of our method approximate rather closely the upper bound of the level of agreement between the experts. Further, we have shown the merit of promoting shared cross-subset patterns and neutralizing topic-specific regularities in a newly introduced dedicated computational step. Methods that do not consider this direction (IB) or that incorporate it within a more conventional cost based search (IB-SI) yield notably poorer performance.

Table 1: Mean Jaccard scores for several methods, examined over of the 16 religion-pair evaluation cases (incorporating mean Jaccard scores over 2–16 clustering configurations, see text). The differences between most CP variants and cross-expert agreement are not statistically significant. The differences between IB, IB-SI and CP_{III} with $\eta = 0.83$ and expert agreement are significant (two-tailed *t*-test, df = 15, $p < 0.01$).

	$\eta = 0.48$	$\eta = 0.56$	$\eta = 0.67$	$\eta = 0.83$
CP	0.405	0.383	0.400	0.394
CP_I	0.416	0.400	0.415	0.399
CP_{II}	0. 410	0.387	0.409	0.417
CP_{III}	0. 405	0.420	0.370	0.293
IB: 0.1734		IB-SI ($\gamma = 0.07$): 0.1995		
Agreement between the experts: 0.462				

6 Disscussion

In this paper, we studied and demonstrated the cross partition method, a computational framework that addresses the task of identifying analogies and correspondences in texts. Our approach to this problem bridges between cognitive observations regarding analogy making, which have inspired it, and unsupervised learning techniques.

While previous cognitively-motivated computational frameworks required structured input (e.g. Falkenhainer et al., 1989), the CP method adapts distributional clustering (Pereira et al., 1993), a standard approach applicable to unstructured data. Unlike standard clustering, the CP method considers an additional source of information: pre-partition of the clustered data to several topical subsets (originated in different sub-corpora) between which a correspondence is drawn.

The innovative aspect of the cross-partition method lies in distinguishing feature information that cuts across the given pre-partition to subsets

versus subset-specific information. In order to incorporate this aspect within distributional clustering, the CP method interleaves several update steps, each locally optimizing a different cost term.

Our experiments demonstrate that the CP method is capable of revealing interesting and non-trivial corresponding themes in texts. The results obtained with most variants of the CP method, with suitable tuning of the parameters, outperform comparable methods – standard distributional clustering and the IB-SI method – and are rather close to the level of agreement between experts.

The CP method revealed, along various resolution levels, meaningful themes that to our understanding can be considered prominent constituents of Religion. It would be an interesting challenge to apply the CP framework further for other tasks, possibly with more practical flavor, such as comparing and detecting commonalities between commercial products and firms, identifying equivalencies and precedents in legal cases and so on.

References

Gal Chechik and Naftali Tishby. 2003. Extracting relevant structures with side information. In S. Becker, S. Thrun, and K. Obermayer (eds.), *Advances in Neural Processing Information Systems 15 (NIPS 2002)*, pp. 857-864.

Ido Dagan, Zvika Marx and Eli Shamir. 2002. Cross-dataset clustering: Revealing corresponding themes across multiple corpora. In D. Roth and A. van den Bosch (eds.), *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, pp. 15-21.

Brian Falkenhainer, Kenneth D. Forbus and Dedre Gentner. 1989. The structure mapping engine: Algorithm and example. *Artificial Intelligence*, 41(1):1-63.

Tomas Gedeon, Albert E. Parker, and Alexander G. Dimitrov, 2003. Information distortion and neural coding. *Canadian Applied Mathematics Quarterly* 10(1):33-70.

Hang Li. 2002. Word Clustering and Disambiguation based on co-occurrence data, *Natural Language Engineering*, 8(1):25-42.

Zvika Marx, Ido Dagan and Eli Shamir. 2004. Identifying structure across pre-partitioned data. In S. Thrun, L. Saul, and B. Schölkopf (eds.), *Advances in Neural Information Processing Systems 16 (NIPS 2003)*, pp. 489-496.

Dekang Lin and Patrick Pantel. 2002. Concept Discovery from Text. In *Proceedings of Conference on Computational Linguistics (COLING-02)*, pp. 577-583.

Fernando C. Pereira, Nftali Tishby and L. J. Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics ACL '93*, pp. 183-190.

Ninian Smart. 1996. *Dimensions of the Sacred: An Anatomy of the World's Beliefs*. University of California Press, Berkeley and Los Angeles, CA.

Naftali Tishby, Fernando C. Pereira and William Bialek. 1999. The information bottleneck method. In *37th Annual Allerton Conference on Communication, Control, and Computing*, pp. 368-379.

Appendix

This appendix specifies the four "local" cost terms mentioned in Section 4 and describes how the CP algorithmic framework (Fig. 2) modifies them.

Step (1) of the CP framework computes $p(c|x)$ values that reduce the value of the following term:

$$F^{CP1} \equiv \underline{H(C)} - H(C|X) + \beta \hat{H}^*(Y|C),$$

where $\hat{H}^*(Y|C) \equiv -\sum_x p(x) \sum_c p(c|x) \sum_y p(y|x) \log p^*(y|c)$. The $p^*(y|c)$ values are considered as if they are constant.

Step (2) computes $p(c|x)$ values reducing the value of

$$F^{CP2} \equiv -\sum_x p(x) \sum_c p(c|x) \sum_y p(y|x) \sum_w p(w|x) \log p(y|c,w),$$

which is equal to $H(Y|C,W)$, subject to an independence assumption extending the assumption explicated in footnote 4, namely for each feature y, cluster c, and pregiven subset w: $p(c,y,w) = \sum_x p(x)\, p(c|x)\, p(y|x)\, p(w|x)$.

Step (3) finds $p^*(c|y)$ values that reduce the value of

$$F^{CP3} \equiv \underline{H^*(C)} - H^*(C|Y) + \eta \hat{H}(Y|C,W),$$

where $H^*(C|Y) = -\sum_y p(y) \sum_c p^*(c|y) \log p^*(c|y)$ and $\hat{H}(Y|C,W) \equiv -\sum_w p(w) \sum_y p(y) \sum_c p^*(c|y) \log p(y|c,w)$, which is equal to the conditioned entropy $H(Y|C,W)$ under an assumption that W is independent of C and Y. The $p(y|c,w)$ values in this term are considered as if they are held constant.

Step (4) finds $p^*(y|c)$ values that reduce the value of

$$F^{CP4} = -\sum_y p(y) \sum_c p^*(c|y) \log p^*(y|c),$$

which can be denoted $H^*(Y|C)$. The $p^*(c|y)$ values are considered as if they are constant.

The underlined $\underline{H(C)}$ and $\underline{H^*(C)}$ terms in F^{CP1} and F^{CP3} are optional; there inclusion implies the inclusion of the prior terms in steps (1) and (3) of the algorithm (see Figure 2).