Discovering Implicit Meanings of Cultural Motifs from Text

Anurag Acharya[1,2] **Diego Casto Estrada**[1] **Shreeja Dahal**[1]
W. Victor H. Yarlott[1] **Diana Gomez**[1] **Mark A. Finlayson**[1]

[1] Florida International University, Miami, FL
markaf@fiu.edu

[2] Pacific Northwest National Laboratory, Richland, WA
anurag.acharya@pnnl.gov

Abstract

Motifs are distinctive, recurring, widely used idiom-like words or phrases, often originating in folklore and usually strongly anchored to a particular cultural or national group. Motifs are significant communicative devices across a wide range of media—including news, literature, and propaganda—because they can concisely imply a large set of culturally relevant associations. One difficulty of understanding motifs is that their meaning is usually implicit, so for an out-group person the meaning is inaccessible. We present the Motif Implicit Meaning Extractor (MIME), a proof-of-concept system designed to automatically identify a motif's implicit meaning, as evidenced by textual uses of the motif across a large set data. MIME uses several sources (including motif indices, Wikipedia pages on the motifs, explicit explanations of motifs from in-group informants, and news/social media posts where the motif is used) and can generate a structured report of information about a motif understandable to an out-group person. In addition to a variety of examples and information drawn from structured sources, the report includes implicit information about a motif such as the type of reference (e.g., a person, an organization, etc.), it's general connotation (strongly negative, slightly negative, neutral, etc.), and it's associations (typically adjectives). We describe how MIME works and demonstrate its operation on a small set of manually curated motifs. We perform a qualitative evaluation of the output, and assess the difficulty of the problem, showing that explicit motif information provided by cultural informants is critical to high quality output, although mining motif usages in news and social media provides useful additional depth. A system such as MIME, appropriately scaled up, would potentially be quite useful to an out-group person trying to understand in-group usages of motifs, and has wide potential applications in domains such as literary criticism, cultural heritage, marketed and branding, and intelligence analysis.

1 Introduction

Motifs can be simply described as recurring cultural "memes" that are grounded in stories. Motifs often originate in folklore, but are ubiquitous and can be found anywhere that language is influenced by culture. They are interesting and useful because they provide a compact source of cultural information: they concisely communicate a constellation of related cultural ideas, associations, assumptions, and knowledge. One common western motif that illustrates the importance and information density of motifs is *troll under the bridge*. One folktale containing the motif, *The Three Billy Goats Gruff*, involves a troll, hiding under a bridge, who tries to devour the goats as they attempt to cross. The motif is found across the folklore of Northern Europe, especially Norway. To members of many western cultures, invoking this motif brings a number of related ideas to mind that are not directly communicated by the surface meaning of the words: the bridge is along the critical path of the heroes and they must cross it to achieve their goal; the troll lives under the bridge, surprising those who attempt to cross it; the troll tries to kill, eat, or otherwise extract some value from the would-be crossers; the troll is a squatter, not the officially sanctioned master of the bridge; and the troll usually meets his end at the hands the hero. The modern utility of the motif as a communicative device is clearly visible in the common term *patent troll*, a person or organization that claims illegitimate ownership over ideas and attempts to extract value from companies who have related products related to those ideas. Here we see an analogical transfer of cultural attributes from the troll of folklore to the "troll" of patents.

Although the example above is drawn from folklore, motifs have importance beyond folktales: they occur in modern stories, news articles, opinion pieces, press releases, propaganda, novels, movies, plays—indeed, anywhere that culture impinges on

Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 46–56

language. One powerful modern example is the use of the *Pharaoh* motif in modern Middle Eastern discourse. The Pharaoh, which refers to the Pharaoh who opposes Moses in the narrative found in Qu'ran, is an arrogant and obstinate tyrant who oppresses the chosen people, defies the will of God, and is punished for it. In modern Islamist extremist discourse, the term *Pharaoh* has been invoked against leaders such as Anwar Sadat of Egypt, Ariel Sharon of Israel, and George W. Bush, the last of whom Osama bin Laden referred to as the "pharaoh of the century" (Halverson et al., 2011). Without understanding the implications of the *Pharaoh* motif, we would be unable to understand both the content of this message (that these leaders are being cast as oppressors) and the cultural group for whom this message was intended (the chosen people, in this case the Ummah).

Work in cognitive psychology has shown that in-group members understand the associations for motifs from their group better than out-group persons (Acharya, 2022). Critically, people can recognize that a motif is being used, regardless of their group membership, and out-group speakers recognize that those motifs carry special meaning and are aware that they're missing out on that meaning. A natural next step, then, is some way of making motif meanings available to out-group persons.

Unfortunately, current resources describing motifs are not sufficient for out-group people to understand the deeper meaning and context of a motif. Let us take an example of a Hindu motif, *Saraswati*. If one looks up *Saraswati* on Google or Wikipedia, one can read that she is the goddess of knowledge, wisdom, music and art, among other things. But this gives no insight into the meaning of the common phrase "Kartik has Saraswati on his tongue." It might sound like that means Kartik is highly knowledgeable or wise, but what it actually means is Kartik never tells a lie. Existing resources, like motif indices, dictionary definitions, and encyclopedia entries do not usually bring forth the nuance and the gravity of meaning that motifs typically carry. A number of additional sources of information are needed, including example usages, to provide the larger context of motif meaning.

To address this problem, we propose the *Motif Implicit Meaning Extractor* (MIME), a proof-of-concept system that can extract and organize various implicit meaning of motifs for presentation to an out-group audience for better understanding. The MIME takes as a query a chunk of text that contains a motif of interest, and uses information from various sources like encyclopedia entries, explanations by in-group informants, and online discourse (e.g., news or social media) to generate a structured report on the common meaning and usage of the motif. Appendix A shows a sample MIME report.

The paper is organized as follows. We first review related work on motifs and their computational processing (§2). Then we list the kinds of implicit information that we would like to extract (§3, followed by the data we collected to support this extraction (§4). Next we explain in detail the system architecture (§5). We then describe our qualitative evaluation, showing that critical information presented in the reports is more likely present in explicit in-group explanations and implicit usage, rather than in Wikipedia entries or news usages (§5.5). Finally, we discuss limitations of this proof-of-concept and map out future work (§7).

2 Related Work

2.1 Motifs

Stith Thompson informally defined a motif as an item "worthy of note because of something out of the ordinary, something of sufficiently striking character to become a part of tradition, oral or literary. Commonplace experiences, such as eating and sleeping, are not traditional in this sense. But they may become so by having attached to them something remarkable or worthy of remembering." (Thompson, 1960, p. 19). In folklore, motifs are preferentially retained throughout retellings and recombinations of tales due to their striking nature and the density of information they communicate. Folklorists have long hypothesized that a tale's specific composition of motifs can be used to trace the tale's lineage (Thompson, 1977, Part 4, Chapter V). This has led folklorists to construct motif indices that identify motifs and note their presence in specific tales (usually as represented in a particular folkloristic collection). The most well-known motif index is the Thompson motif index (Thompson, 1960). Thompson's index designates each motif with a code; for example, *troll under a bridge* is referenced by the codes G304 and G475.2. In this case, *troll under a bridge* is represented by two motifs as Thompson generalizes trolls to ogres, a general class of monstrous beings; thus, the motifs are *troll as ogre* (G304) and *ogre attacks intruders on bridge* (G475.2). Thompson

noted that motifs generally fall into one of three subcategories (Thompson, 1977, pp. 415–416): events, characters, or props. Examples of these include *Hero rescuing a Princess* (B11.11.4; event), *Old Man Coyote* (A177.1, character), and *Magic Carpet* (D1155, prop).

Folklorists were the originators of the idea of motif, and constructed motif indices that identify motifs and note their presence in specific folktales. As stated previously, the most well-known motif index is Thompson's motif index (TMI) (Thompson, 1960), which integrates a number of prior indices and references folktales from over 600 collections, indexed to 46,248 motifs and submotifs. In addition to this, Thompson provides substantial discussion on motifs and the compilation of motif indices in his book *The Folktale* (Thompson, 1977). Additionally, there are many motif indices which target specific cultures and periods, for example, early Irish literature (Cross, 1952), traditional Polynesian narratives(Kirtley, 1971), Japanese folk-literature(Ikeda, 1971), or early Icelandic literature(Boberg, 1966). While Thompson's motif index is perhaps the primary source of motif information used today, it has been criticized because of overlapping motif subcategories, censorship (primarily of obscenity), and missing motifs (Dundes, 1997).

2.2 Computational Approaches to Motifs

Darányi (2010) called attention to the need for research into the automation of extraction and annotation of motifs in folklore, and suggested that motifs have application in storing, indexing, and retrieving documents based on the motifs contained within. Work has also been done examining the shortcomings and potential applications of motifs. For example, Darányi and Forró (2012) determined, based on cluster analysis, that motifs may not be the highest level of abstraction in narrative, echoing criticisms that many motifs are interdependent (Dundes, 1997). Darányi et al. (2012) have made substantial headway towards using motifs as sequences of "narrative DNA", and Ofek et al. (2013) have demonstrated learning tale types based on these sequences. Declerck et al. (2012) have also done work on converting electronic representations of TMI and ATU (Uther, 2004) to a format that enables multilingual, content-level indexing of folktale texts, building upon past work (Declerck and Lendvai, 2011). Currently, this work appears to be focused on the descriptions of motifs and tale types, without reference to the stories.

2.3 Relation Detection, Information Extraction, and Template Filling

The MIME can be thought of as a mix of targeted information extraction, relationship extraction, and template filling, all tasks that have seen much attention in the past.

Soares et al. (2019) used distributional similarity to build a general relation extraction system using BERT (Devlin et al., 2018). Similarly, Wu and He (2019) also used BERT to perform relation classification while also incorporating information about large entities. Meanwhile Ye and Ling (2019) and Wang et al. (2016) implement CNN-based few-shot relation classifiers. There are several other systems that perform well on relation detection, such as Bastos et al. (2021), a neural network based model; Kim et al. (2019), an RNN based model; and Cai et al. (2016), a BRCNN based model.

There has also been much work on information extraction (IE), both in general and specific types of relations. One important work is OpenIE by Stanovsky et al. (2018), a system which re-frames open information extraction as a sequence tagging problem. Similar systems include those by Cetto et al. (2018), Gashteovski et al. (2017), and Bhutani et al. (2016) which all perform close to the state of the art. Importantly, OpenIE defined a standard relation set, and certain systems have focused on improving perform on specific relations in that set. For example, Pal et al. (2016) focuses on nominal OpenIE, which finds an efficient way to extract open relations for compound noun phrases. Similarly, Saha et al. (2017) focuses specifically on numerical relations to extract OpenIE tuples, and Saha et al. (2018) addresses the issue of extracting relation tuples for conjunctive sentences.

It is worth noting that we did try to see if our task could be achieved by using an existing off-the-shelf IE systems. We tested several state-of-the-art IE systems including OpenIE (Stanovsky et al., 2018), CALMIE, (Saha et al., 2018) and BONIE (Saha et al., 2017). We found the output of these systems were not adequate to the task, which we traced to the fact that that most IE systems focus on each sentence in isolation, i.e., without context. Taking advantage of context required breaking out individual IE subtasks in a different way, which led to the MIME approach described in Section 5.

Finally, a number template filling systems are related to the MIME task. For example, Jean-Louis et al. (2011) combines text segmentation and graph techniques to perform template filling. Another is Miliani et al. (2019) which splits text into frames in order to accomplish slot filling. But perhaps the work that most closely relates to this work is work done by Chambers and Jurafsky (2011) which combines the tasks of information extraction and template filling, but is able to do so without having a fixed template for the output in advance. For this work, we draw on all these tasks, mixing and matching our implementations to maximize explainability of the proof-of-concept system. In particular, we defer the use of black-box neural systems, which can be used to optimize performance, for later implementations.

3 Types of Implicit Information

There are several categories of implicit information, listed below, we would like to expose about motifs to improve understanding by out-group persons.

3.1 Type of Reference

The first piece of information that is useful to understand a motific usage is the answer to the question "What kind of object does this motif refer to?" Thompson roughly split motifs into three broad types: events, characters, or props. Which one a motif refers to may not always be obvious from context. For example, "That person is such an *Amalek*." makes clear that the object being referred to is a person, but the usage "Sorry, Jack: They are bad, but they are no *Amalek*." or even "What an *Amalek*." is ambiguous.

3.2 General Connotation

The second piece of implicit information that is useful is the general positive, neutral, or negative connotation of the motif, which may unknown or be different than commonly assumed. In the "That person is such an *Amalek*." example above, the connotation alone can give us significant insight into the meaning. *Amalek* refers to an ancient ruler in the Hebrew Bible who was a well-known persecutor of the Jews, and is usually used to speak negatively about a person. Another good example is the motif *Leprechaun*: while in the broader Western culture it is assumed to have a general positive connotation, within an Irish context specifically it is more neutral, being a mix of positive and negative elements, as Leprechauns are mischievous, tricky beings, but not necessarily evil.

3.3 Associations

Finally, a deeper understanding of a motif involves understanding the specific associations or implications it calls to mind. With the example above of *troll under the bridge* in the introduction, there are a variety of implications about the motives and legitimacy of the troll that the motif calls to mind. As another example, in the case of *Finn McCool*, using this motif to describe a person implies potentially that they are incredibly smart or powerful, or act as a savior of many people or protector of the land.

4 Data

4.1 Motif Selection

There are literally tens of thousands of motifs listed within the many hundreds of motif indices that have been written since the late 1800s. Most of these motifs, furthermore, are not actually well known or commonly used within the relevant group. This sparsity presents a challenge for identifying motifs that are actually used in communicative context of interest. It was not our goal in this work to solve this particular problem. Instead, our goal was to develop a method that, given a list of motifs known to be in use, could extract the relevant implicit information. Therefore, we first had to manually select a set of target motifs with which we could develop our proof-of-concept system.

We determined three criteria for selecting motifs on which to demonstrate the MIME. First, they needed to be "high quality" motifs from a clearly identifiable group, which we define below. Second, we needed access to members of the relevant group to gather in-group explanations of motif meaning, to be used both as system data and evaluation materials. Finally, the number of motifs needed to be small enough to be manageable within the project scope and budget. We defined "high quality" motifs according to the following three characteristics:

1. **Common Use**: Selected motifs should be in common use in modern communicative contexts. The simplest test of this was a keyword search to see if the actual words expressing the motif were used either on social media, such as Twitter, or in the news. This criteria simplified the search—if we couldn't find it, we did not include it.

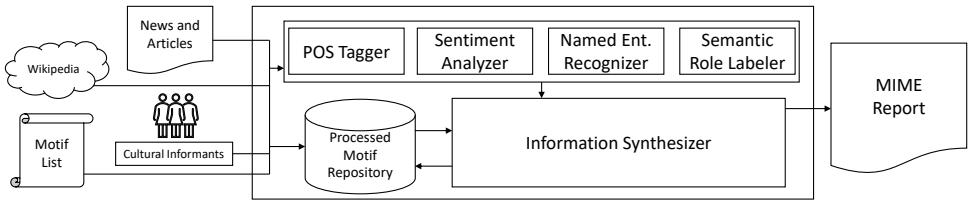

Figure 1: The main architecture design of the Motif Implicit Meaning Extractor (MIME), including the external components that feed into the system.

2. **Clear Source**: Selected motifs must have a clearly identifiable source within the group. By *source*, we mean an associated, well-known story within the same body of folklore as other motifs for the group. This criteria is intended to provide proof of relevance for the motif to the group in question. If a motif had no definitive source within the folklore of the cultural group we excluded it.

3. **Strength**: Selected Motifs should be preferentially used in a motific way, meaning their usage usually draws on the implicit meaning, rather than being used literally or eponymically (i.e., as a name). This was a subjective judgement of *how* the motifs were used when we found them in news.

Intersecting these three criteria led us to select roughly ten motifs from each of three groups: Irish, Jewish, and Puerto Rican. These were groups that had easily accessible motif indices, quite of a bit of activity by group members on news and social media, and whose members were accessible to us through various connections and contacts. Regarding the motif indices, for Irish, we used T.P. Cross's *Motif-Index of Early Irish Literature* (Cross, 1952) as a main source; for Puerto Rican, we drew motifs from S.R. Lamarche's *The Mythology and Religion of the Tainos* (Hurley et al., 2021), R.E. Alegría's *The Three Wishes: A Collection of Puerto Rican Folktales* (Alegria et al., 1969), and J. Ramírez-Rivera's *Puerto Rican Tales: Legends of Spanish Colonial Times* (Ramírez-Rivera et al., 1977); and for Jewish motifs, we referenced D.N. Noy's *Motif-index of Talmudic-Midrashic literature* (Noy, 1954). We read through these indices, and in discussion with our in-group informants and investigation on social media and news search engines, we settled on the following set:

Irish (13) The Salmon of Wisdom, Finn McCool, Leprechaun, King Conchobar, Aos Si, Banshee, Cu Chulainn, the Wren, the Magic Harp, Tir Na Nog, Shamrock, Fairy Fort, the Children of Lir

Jewish (9) Haman, Golem, Amalek, Babel, Leviathan/Behemoth, 70 Languages, Name in Vain, the Ark of the Covenant, Kiddush Hashem

Puerto Rican (12) Reyes Magos/Three Kings, Agueybana, Atabey, Roberto Cofresi, Divina Providencia, Guanina, Juan Bobo, Yocahu, the Coqui, Hormigueros, Jibaro/Jibarito, Chupacabra

4.2 Wikipedia Articles

Many motifs have corresponding Wikipedia articles which contain some useful information (although, as noted above, are not by themselves sufficient to understand motif usages). For each motif in our set we found and downloaded the relevant Wikipedia article if it existed (19 out of 34). We converted the article to plain text, eliminating formatting and surrounding elements (e.g., links to other pages, donation advertisements). We further eliminated irrelevant article sections of the articles (e.g., References, History), while retaining sections like Background Information, Summary, and Traditions. Although we only looked at Wikipedia in this work, it would be possible to expand this portion of the data by adding other encyclopedias, thesauri, or dictionaries.

4.3 In-Group Informant Explanations

For each motif we also interviewed our in-group informants as to their meaning. In these short interviews we provided some example usages of the motif (found via keyword search) and asked our informants to explain the meaning of the motif in context. We produced short descriptions that summarized the answers of multiple informants, which were then verified by the informants themselves. Examples for three motifs are as follows:

Finn McCool (Irish, Character) Incredibly smart, powerful savior of many people. Considered a protector of the land.

Kiddush (Jewish, Event) A prayer; more commonly invoked as "kiddush hashem" which is representing God, the Torah, and the Jewish people in the best light possible.

Coqui (Puerto Rican, Prop) Small frog species native to Puerto Rico. Known for being the national symbol of Puerto Rico.

4.4 Online News

We also collected a large number of online news articles where the motifs were used in a motific sense. We obtained English texts through NexisUni[1], a university version of LexisNexis, a tool for searching through news articles, which provides worldwide scope for news and related text. We searched for motif terms and batch downloaded these articles, as allowed by the University's license. These articles were then further processed by a Lucene-based lexical matcher, with fuzzy rules for a variety of lexical forms for each motif, to verify the presence of motif terms (called *motif candidates*) for inspection by our annotators. This dataset comprises 26,078 motifs candidates across 7,955 texts.

We hired annotators who identified as members of the groups in question, which was determined through an interview. We also required annotators to possess a college degree and be fluent in English. We hired six annotators total (two annotators per group) to perform the double-blind annotations. Annotators were asked, for each motif candidate, to identify whether or not the usage invoked the implicit meaning of a motif (e.g., referring to something large and monstrous as a *behemoth*). A motif candidate that invoked the in-group meaning was called a *motif instance*. The Jewish and Puerto Rican teams participated produced motif instance annotations with an average agreement of $\kappa > 0.7$ while the Irish team produced annotations with an average agreement of $\kappa > 0.55$. The annotation resulted in 1,723 motif instances (159 Irish, 1,215 Jewish, and 349 Puerto Rican).

5 Computational Pipeline

We implemented a traditional NLP pipeline, rather than an end-to-end neural system (generative or otherwise) for several reasons. First, a pipelined system is more explainable, allow us to do error analysis as to whether individual pieces of information were easier or harder to extract, allowing us to refine our task definition. Second, we have relatively little data with which to fine-tune a neural system. In particular, due to the difficulty of collecting high-quality information about specific motifs, (a) we have only 34 motifs in our demonstration, (b) we have only 1,723 annotated motif instances (159 Irish, 1,215 Jewish, and 349 Puerto Rican), and (c) the in-group informant explanations are only a few sentences long each. Third, since we are concentrating on the structure of the task and not optimizing for performance, a pipelined system is more useful; once the task is well defined and more data has been collected, later work can focus on integrating neural architectures for optimization. Finally, our anecdotal experiments with the latest generative text models (e.g., chatGPT) found that it was quite poor in its ability to generate accurate information about motifs.

The proof-of-concept MIME system has four components, described below. We applied all components to all of the data, including the Wikipedia articles, in-group informant explanations, and 100-token windows surrounding annotated motif instances in the online news. Information extracted from these data are stored in a database indexed to each motif, and can be used to generate a motif report (as shown in Appendix A).

5.1 Part-of-Speech Tagger

Part-of-Speech (POS) tagging is an important piece of information for understanding type of reference and associations. We used SpaCy's (Honnibal et al., 2020) medium-sized CPU-optimized POS tagger.

5.2 Named Entity Recognizer

The types of Named Entities (NEs) mentioned in relation to a motif again speaks to type of reference and associations. We used an implementation of the baseline Named Entity Recognizer (NER) model from (Peters et al., 2017), which was provided by AllenNLP in the form of its ELMo-Based NER. This model achieves a reported score of 96% F1 score on the CoNLL-2003 validation set (Tjong Kim Sang and De Meulder, 2003). The AllenNLP NER uses twenty-one classes, organized into a hierarchy. We reduce the number of possible NER tags for the motifs by consolidating named entities tags into their top-level parent class. This resulted in

[1] https://www.lexisnexis.com/en-us/professional/academic/nexis-uni.page

just four classes: Person, Organization, Location, and Miscellaneous.

5.3 Sentiment Analyzer

Sentiment is critical to understanding the implicit connotation of a motif. We used AllenNLP's sentiment analyzer, which outputs either a 0 or a 1 using an LSTM classifier with GloVe embeddings (Pennington et al., 2014) with a reported performance of 87% accuracy on the Stanford Sentiment Treebank corpus (Socher et al., 2013), which is near to state-of-the-art performance.

5.4 Semantic Role Labeler

Semantic role labeling exposes subject-relation-object structures within the texts, which is relevant to the type of reference and implicit associations. We used AllenNLP's Semantic Role Labeler (SRL), which is a slightly modified version of a BERT based model (Shi and Lin, 2019), with a reported F_1 of 86.49 on the Ontonotes 5.0 (Weischedel et al., 2013). While AllenNLP's SRL generates role labels in the full PropBank scheme (Babko-Malaya, 2005), this level of complexity was not necessary for our proof-of-concept system. Therefore we mapped the various SRL tags into just three main roles: ARG0 (agent), ARG1 (patient), and ARG2 (location/instrument/beneficiary/etc.), while ignoring all other roles as "non-specific".

5.5 Information Synthesizer

Information generated by the prior four components over all the data is integrated together in the information synthesizer modules. First, NER taggings are examined from across the all data to determine the most common type reference, which falls into one of the four categories of *Person*, *Organization*, *Location*, and *Miscellaneous*.

Second, the sentiment values across all the data for each motif are averaged and transformed into one of four broad categories, as shown in Table 1.

Sentiment Label	Value
Negative	0 – 0.25
Slightly Negative to Neutral	0.25 – 0.5
Neutral to Slightly Positive	0.5 – 0.75
Positive	0.75 – 1

Table 1: New sentiment categories for the motifs

Third, the SRL Subject-Verb-Object triples for each motif are processed. When the subject or object contains the motif of interest, we look at the other role, and if this is an adjective, this is listed as an association of that motif. If the other role does not contain an adjective, we return to the original sentence to see if there are any adjectives in the sentence within a 21-word window centered on the motif, and include these as associations.

1. **Motif Type**: Whether a motif is event, character, or prop
2. **Origin Culture**: What culture the motif originated from
3. **Usually referred to**: What type of entity the motif refers to, and how it's used in a sentence
4. **Major Associations**: What are the major associations of the motif
5. **General Usage Connotation**: How is the motif typically perceived?
6. **Examples**: Example usages of the motif
7. **Background**: Short explanation of the motif

Table 2: Templates to be filled out for the MIME report

With all the above information processed from the data and included in the MIME's database, the MIME can produce a report for each motif of interest. We treated report generation as a specialized template-filling task. The template fields are shown in Table 2. To match the fields in the template, all the information in the synthesizer goes through the process of converting tags and stubs into phrases and/or sentences and merging information together from different modules to put relevant information together. One possible source of error in this process is if the information from the external sources or the catalog is too long. To avoid this, we truncate the information to a fixed length. The final result of this process is a filled out template, which is the final MIME Report, as shown in Appendix A.

6 Results and Discussion

Since this task is a version of targeted Information Extraction and the output is a human-readable report, it is challenging to apply traditional evaluation metrics to assess effectiveness. We performed a qualitative analysis of the reports, comparing them against the explanations provided by the in-group informants as well as our own general understanding of the meaning of motifs.

In general, MIME did well at finding implicit information for motifs that have several instances of proper usage in news. It was able to identify the type of reference of the motifs fairly accurately (only 1 answer incorrect/missing, because of insufficient data), which is useful because this information was not provided by the in-group explanations. MIME was also able to reveal the general

connotation of the motifs correctly across most of the motifs (6 incorrect). In some cases, like *Leprechaun*, from usages in the news data it was able to infer a connotation closer to that provided by the in-group explanations, despite the popular understood meaning being slightly different. This shows that the MIME approach adds value beyond what is provided in the explicit in-group explanations.

We also evaluated the the proportion of information derived from our three sources. We examined the three main types of implicit information—type of reference, general connotation, and associations—and counted how often that information could be traced back to a specific type of source (Wikipedia, Explanations, or Online News). The summary of these results are shown in Tables 3 and 4. The results show that the bulk of the associations (roughly 80%) come from the in-group explanations, with approximately 14% coming from news. This result shows the importance of the in-group explanations, and that the meaning of motifs, in general, is not discoverable from usage in modern news alone. While the Wikipedia text was not particularly information dense for this task, in our qualitative analysis it played a crucial role of providing additional context in terms of clear-to-read summaries and background, thereby providing some level of utility. This is especially visible for the Puerto Rican motifs, where the Wikipedia data was especially devoid of depth, and the reports are notably less rich than the other two groups. Overall, we see that the results show the potential of this approach, consistent with its proof-of-concept stage of development.

	Wiki	Explanations	News
Reference	1.4%	3.4%	95.2%
Connotation	0.7%	2.2%	97.7%
Association	2.4%	79.5%	13.7%
Overall	1.5%	28.4%	68.9%

Table 3: Proportion of implicit information that can be attributed to each source, grouped by information type.

	Wiki	Explanations	News
Irish	2.7%	30.6%	68.0%
Jewish	1.8%	22.9%	73.1%
PR	0%	31.6%	65.6%
Overall	1.5%	28.4%	68.9%

Table 4: Proportion of implicit information that can be attributed to each source, by group

7 Limitations and Future Work

As a proof-of-concept, MIME has a number of important limitations. First, by design the system does not attempt to extract information that must be obtained via multiple steps of inference. Second, the system sometimes reports associations that are inconsistent with those reported by the cultural informants. In our qualitative evaluation this is mostly because the motif seems to have been used incorrectly by out-group persons in online news. A few of these mistakes can be attributed to an adjective that is near the motif but was grammatically attached elsewhere. A more sophisticated approach to attribute extraction could mitigate this issue, or collecting further example usages from cultural informants and assigning higher weight to those. Finally, if the data sources lack relevant information the system will naturally fall short; this was evident in particular for the Puerto Rican motif *Yocahu*, where there were very few news usages.

One key future step is to look more comprehensively at the discourse context of the motifs instances. For instance, we not only have sentiment values and linguistic properties for the sentences with motifs themselves, but also for other sentences that form part of the discourse that involves the motif. Using this discourse relation information, we will be able to more accurately predict the associations for the motifs. We could not implement such an approach because of lack of sufficient examples of motifs in narrative text; with more data of real-world narrative usage of motifs, this could potentially be implemented in the future.

The importance of the in-group explanations for implicit associations is also revealing. Looking only at usages in modern discourse would make it very hard to extract these associations; on the other hand, asking in-group persons to explain every motif is time-consuming and laborious. Here, one untapped source of information would be the original narratives themslves, from which the original meaning of the motifs are derived. Ideally, a next step would be to mine this information for the original context and meaning of the motifs.

One main limitation of MIME, of course, is that it requires manual identification and annotation of the motifs. Recent work by Yarlott et al. (2021) demonstrated a preliminary motif detection system which perhaps could form the foundation for a scaling up of the MIME to as-yet-unanalyzed motifs.

Acknowledgments

This work was supported in part by DARPA contract FA8650-19-C-6017. The primary author of this research, Dr. Acharya, is now at Pacific Northwest National Laboratory, which is operated by Battelle Memorial Institute for the U.S. Department of Energy under contract DE-AC05-76RLO1830, but this work was completed while he was at FIU.

References

Anurag Acharya. 2022. *Integrating Cultural Knowledge into Artificially Intelligent Systems: Human Experiments and Computational Implementations*. Ph.D. thesis, Florida International University.

R.E. Alegria, R.E. Alegría, L. Homar, and E. Culbert. 1969. *The Three Wishes: A Collection of Puerto Rican Folktales*. Harcourt, Brace & World.

Olga Babko-Malaya. 2005. Propbank annotation guidelines.

Anson Bastos, Abhishek Nadgeri, Kuldeep Singh, Isaiah Onando Mulang, Saeedeh Shekarpour, Johannes Hoffart, and Manohar Kaul. 2021. Recon: relation extraction using knowledge graph context in a graph neural network. In *Proceedings of the Web Conference 2021*, pages 1673–1685.

Nikita Bhutani, HV Jagadish, and Dragomir Radev. 2016. Nested propositions in open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 55–64.

Inger Margrethe Boberg. 1966. *Motif-index of early Icelandic literature*. Munksgaard.

Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. Bidirectional recurrent convolutional neural network for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 756–765.

Matthias Cetto, Christina Niklaus, André Freitas, and Siegfried Handschuh. 2018. Graphene: Semantically-linked propositions in open information extraction. *arXiv preprint arXiv:1807.11276*.

Nathanael Chambers and Dan Jurafsky. 2011. Template-based information extraction without the templates. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 976–986.

Tom Peete Cross. 1952. *Motif-index of early Irish literature*. Indiana University.

Sándor Darányi. 2010. Examples of Formulaity in Narratives and Scientific Communication. In *Proceedings of the First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts*, pages 29–35.

Sándor Darányi and László Forró. 2012. Detecting Multiple Motif Co-occurrences in the Aarne-Thompson-Uther Tale Type Catalog: A Preliminary Survey. *Anales de Documentación*, 15(1).

Sándor Darányi, Peter Wittek, and László Forró. 2012. Toward Sequencing "Narrative DNA": Tale Types, Motif Strings and Memetic Pathways. In *Third Workshop on Computational Models of Narrative (CMN)*, pages 2–10, Istanbul, Turkey. European Language Resources Association (ELRA).

Thierry Declerck and Piroska Lendvai. 2011. Linguistic and semantic representation of the thompson's motif-index of folk-literature. In *Research and Advanced Technology for Digital Libraries*, pages 151–158. Springer.

Thierry Declerck, Piroska Lendvai, and Sándor Darányi. 2012. Multilingual and Semantic Extension of Folk Tale Categories. In *Proceedings of the 2012 Digital Humanities Conference (DH 2012)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alan Dundes. 1997. The motif-index and the tale type index: A critique. *Journal of Folklore Research*, pages 195–202.

Kiril Gashteovski, Rainer Gemulla, and Luciano del Corro. 2017. MinIE: Minimizing facts in open information extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2630–2640, Copenhagen, Denmark. Association for Computational Linguistics.

Jeffry R Halverson, Steven R Corman, and HL Goodall Jr. 2011. *Master narratives of Islamist extremism*. Palgrave Macmillan.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. *Online code repository*.

A. Hurley, C.R.R. de Arellano, and S.R. Lamarche. 2021. *The Mythology and Religion of the Tainos*. Independently Published.

Hiroko Ikeda. 1971. *A type and motif index of Japanese folk-literature*. Orient Cultural Service.

Ludovic Jean-Louis, Romaric Besançon, and Olivier Ferret. 2011. Text segmentation and graph-based method for template filling in information extraction. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 723–731.

Byoungjae Kim, KyungTae Chung, Jeongpil Lee, Jungyun Seo, and Myoung-Wan Koo. 2019. A bi-lstm memory network for end-to-end goal-oriented dialog learning. *Computer Speech & Language*, 53:217–230.

Bacil F Kirtley. 1971. *A motif-index of traditional Polynesian narratives*. University of Hawai'i Press.

Martina Miliani, Lucia C Passaro, and Alessandro Lenci. 2019. Text frame detector: Slot filling based on domain knowledge bases. In *CLiC-it*.

Dov Neuman Noy. 1954. *Motif-index of Talmudic-Midrashic literature*. Indiana University.

Nir Ofek, Sándor Darányi, and Lior Rokach. 2013. Linking Motif Sequences with Tale Types by Machine Learning. In *Proceedings of the 4th Workshop on Computational Models of Narrative (CMN'13)*, volume 32, pages 166–182, Hamburg, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Harinder Pal et al. 2016. Demonyms and compound relational nouns in nominal open ie. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*, pages 35–39.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *ACL*.

J. Ramírez-Rivera, B. Klein, and J. Slemko. 1977. *Puerto Rican Tales: Legends of Spanish Colonial Times*. Ediciones Libero.

Swarnadeep Saha, Harinder Pal, et al. 2017. Bootstrapping for numerical open ie. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 317–323.

Swarnadeep Saha et al. 2018. Open information extraction from conjunctive sentences. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2288–2299.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *ArXiv*, abs/1904.05255.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895.

Stith Thompson. 1960. *Motif-index of folk-literature: a classification of narrative elements in folktales, ballads, myths, fables, mediaeval romances, exempla, fabliaux, jest-books and local legends*, volume 4. Indiana University Press.

Stith Thompson. 1977. *The folktale*. Univ of California Press.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Hans-Jörg Uther. 2004. *The types of international folktales: a classification and bibliography, based on the system of Antti Aarne and Stith Thompson*. Suomalainen Tiedeakatemia, Academia Scientiarum Fennica.

Linlin Wang, Zhu Cao, Gerard De Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention cnns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1298–1307.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.

Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2361–2364.

W Victor H Yarlott, Armando Ochoa, Anurag Acharya, Laurel Bobrow, Diego Castro Estrada, Diana Gomez, Joan Zheng, David McDonald, Chris Miller, and Mark Alan Finlayson. 2021. Finding trolls under bridges: Preliminary work on a motif detector. In *Proceedings of the Ninth Annual Conference of Advancements in Cognitive Systems (ACS) 2021*.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Multi-level matching and aggregation network for few-shot relation classification. *arXiv preprint arXiv:1906.06678*.

A Example Motif Implicit Meaning Extractor Report

The figure below shows a MIME report generated by the system, for the motif *"Leprechaun"*. The report has been manually formatted for the sake of clarity, but the contents remain unchanged.

Title	**Content**
Motif	Leprechaun
Found in	I previously talked about the franchise in my usual breakdown format four years ago, but I've never ranked them. We live in a world where there are seven Leprechaun movies. Seven. I just can't comprehend that. They're not even particularly good movies. Of course, I say this and I own the entire series on blu-ray. Hey, they're still fun and Warwick Davis is always entertaining. So let's look at all seven of these movies and see which is the best!
Motif Type	Character
Origin Culture	Irish
Usually referred to	Mostly used to refer to a person. Generally used as no specific role (eg. subject/object) in a sentence.
Major associations	Tricky, grumpy, short, afraid
General Usage Connotation	Neutral to Slightly Negative
Motific Examples	(1) That old miser is a real leprechaun. (2) That leprechaun at the used car lot really got the better of me. (3) I swear I could jump over Ethan, he's a real leprechaun.
Referential Example	A leprechaun is a type of fairy of the aos si in Irish folklore.
Unrelated Example	N/A—most uses are going to be at least somewhat culturally related.
Eponym Example	The photo you see for Leprechaun, Inc. is a 5,000 year old Dolman, or Portal Tomb, built during the Neolithic Period.
Background	A leprechaun (Irish : /leipreachán/luchorpán/) is a diminutive supernatural being in Irish folklore, classed by some as a type of solitary fairy. They are usually depicted as little bearded men, wearing a coat and hat, who partake in mischief. In later times, they have been depicted as shoemakers who have a hidden pot of gold at the end of the rainbow. Leprechaun-like creatures rarely appear in Irish mythology and only became prominent in later folklore. They are usually depicted as little bearded men, wearing a coat and hat, who partake in mischief. In later times, they have been depicted as shoe-makers who have a hidden pot of gold at the end of the rainbow. Leprechaun-like creatures rarely appear in Irish mythology and only became prominent in later folklore