Extracting Cultural Commonsense Knowledge at Scale

Tuan-Phong Nguyen  
Max Planck Institute for Informatics  
Saarbrücken, Germany  
tuanphong@mpi-inf.mpg.de

Aparna Varde  
Montclair State University  
Montclair, New Jersey, USA  
vardea@montclair.edu

Simon Razniewski  
Max Planck Institute for Informatics  
Saarbrücken, Germany  
srazniew@mpi-inf.mpg.de

Gerhard Weikum  
Max Planck Institute for Informatics  
Saarbrücken, Germany  
weikum@mpi-inf.mpg.de

ABSTRACT
Structured knowledge is important for many AI applications. Commonsense knowledge, which is crucial for robust human-centric AI, is covered by a small number of structured knowledge projects. However, they lack knowledge about human traits and behaviors conditioned on socio-cultural contexts, which is crucial for situative AI. This paper presents Candle, an end-to-end methodology for extracting high-quality cultural commonsense knowledge (CCSK) at scale. Candle extracts CCSK assertions from a huge web corpus and organizes them into coherent clusters, for 3 domains of subjects (geography, religion, occupation) and several cultural facets (food, drinks, clothing, traditions, rituals, behaviors). Candle includes judicious techniques for classification-based filtering and scoring of interestingness. Experimental evaluations show the superiority of the Candle CCSK collection over prior works, and an extrinsic use case demonstrates the benefits of CCSK for the GPT-3 language model. Code and data can be accessed at https://candle.mpi-inf.mpg.de/.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/authors(s).

WWW '23, May 1–5, 2023, Austin, TX, USA  
© 2023 Copyright held by the owner/authors(s).  
ACM ISBN 978-1-4503-9416-1/23/04  
https://doi.org/10.1145/3543507.3583535

1 INTRODUCTION

Motivation. Structured knowledge, often stored in knowledge graphs (KGs) [12, 39], is a key asset for many AI applications, including search, question answering, and conversational bots. KGs cover factual knowledge about notable entities such as singers, songs, cities, sports teams, etc. However, even large-scale KGs deployed in practice hardly touch on the dimension of commonsense knowledge (CSK): properties of everyday objects, behaviors of humans, and more. Some projects, such as ConceptNet [36], Atomic [32], and Ascent++ [21] have compiled large sets of CSK assertions, but are solely focused on “universal CSK”: assertions that are agreed upon by almost all people and are thus viewed as “globally true”. What is missing, though, is that CSK must often be viewed in the context of specific social or cultural groups: the world view of a European teenager does not necessarily agree with those of an American business person or a Far-East-Asian middle-aged factory worker. This paper addresses this gap, by automatically compiling CSK that is conditioned on socio-cultural contexts. We refer to this as cultural CSK or CCSK for short. For example, our CCSK collection contains assertions such as:

- **subject:East Asia, facet:food**, Tofu is a major ingredient in many East Asian cuisines, or
- **subject:firefighter, facet:behavior**, Firefighters use ladders to reach fires.

The value of having a KG with this information lies in making AI applications more situative and more robust.

Consider the conversation between a human and the GPT-3 chatbot1 shown in Fig. 1. The GPT-3-based bot, leveraging its huge language model, performs eloquently in this conversation, but completely misses the point that the user is in China, where dragons are viewed positively and espresso is difficult to get. If we prime

![Figure 1: Human-bot conversations without and with CCSK.](https://candle.mpi-inf.mpg.de/)

---

1 Executed at beta.openai.com/playground using the davinci-002 model at temp=0.7.
the bot with CCSK about Far-East-Asian culture, then GPT-3 is enabled to provide culturally situative replies. If primed with CCSK about European views (not shown in Fig. 1), the bot points out that dragons are portrayed as evil monsters but do not exist in reality and recommends a strong cup of coffee.

State of the art. Mainstream KGs do not cover CCSK at all, and major CSK collections like ConceptNet contain only very few culturally contextualized assertions. To the best of our knowledge, the only prior works with data that have specifically addressed the socio-cultural dimension are the projects Quasimodo [30], StereoKG [7], and the work of Acharya et al. [1]. The latter merely contains a few hundred assertions from crowdsourcing. StereoKG uses a specialized way of automatically extracting stereotypes from QA forums and is still small in size, and Quasimodo covers a wide mix of general CSK and a small fraction of culturally relevant assertions. These are the three baselines to which we compare our results.

Language models (LMs) such as BERT [8] or GPT-3 [5] are another form of machine-based CSK, including CCSK, in principle. However, all LM knowledge is in latent form, captured in learned values of billions of parameters. Knowledge cannot be made explicit; we observe it only implicitly through the LM-based outputs in applications. The example of Fig. 1 demonstrates that even large LMs like GPT-3 do not perform well when cultural context matters.

Approach. CCSK is expressed in text form on web pages and social media, but this is often very noisy and difficult to extract. We devised an end-to-end methodology and system, called CANDLe (Extracting Cultural Commonsense Knowledge at Scale), to automatically extract and systematically organize a large collection of CCSK assertions. For scale, we tap into the C4 web crawl [27], a huge collection of web pages. This provides an opportunity to construct a sizable CCSK collection, but also a challenge in terms of scale and noise.

The output of CANDLe is a set of 1.1M CCSK assertions, organized into 60K coherent clusters. The set is organized by 3 domains of interest – geography, religion, occupation – with a total of 386 instances, referred to as subjects (or cultural groups). Per subject, the assertions cover 5 facets of culture: food, drinks, clothing, rituals, traditions (for geography and religion) or behaviors (for occupations). In addition, we also annotate each assertion with its salient concepts. Examples for the computed CCSK are shown in Fig. 2.

CANDLe operates in 6 steps. First and second, we identify candidate assertions using simple techniques for subject detection (named entity recognition - NER, and string matching), and generic rule-based filtering. Third, we classify assertions into specific cultural facets, which is challenging because we have several combinations of cultural groups and cultural facets, making it very expensive to create specialized training data. Instead, we creatively leverage LMs pre-trained on the Natural Language Inference (NLI) task to perform zero-shot classification on our data, with judicious techniques to enhance the accuracy. Fourth, we use state-of-the-art techniques for assertion clustering, and fifth a simple but effective method to extract concepts in assertions. Lastly, we combine several features to score the interestingness of assertions, such as frequency, specificity, distinctiveness. This way, we steer away from overly generic assertions (which LMs like GPT-3 tend to generate) and favor assertions that set their subjects apart from others.

Table 1: Example assertions of CANDLe, with subjects (cultural groups) of cultural domains, facets and concepts.

| Geography     | Germany   | Drinking   |
|---------------|-----------|------------|
| German beer festivals in October are a celebration of beer drinking. |

| Geography     | East Asia | Food       |
|---------------|-----------|------------|
| Tofu is a major ingredient in many East Asian cuisines. |

| Geography     | South Asia | Traditions |
|---------------|------------|------------|
| In South Asia, henna is often used in bridal makeup or to celebrate festivals. |

| Occupation   | Lawyer    | Clothing  |
|--------------|-----------|-----------|
| Lawyers wear suits to look professional. |

| Occupation   | Firefighter | Behaviors |
|--------------|-------------|-----------|
| Firefighters run into burning buildings to save lives. |

Figure 2: Example assertions of CANDLe, with subjects (cultural groups) of cultural domains, facets and concepts.

Contributions. The main contributions of this work are:

1. An end-to-end methodology to extract high-quality CCSK from very large text corpora.
2. New techniques for judiciously classifying and filtering CCSK-relevant text snippets, and for scoring assertions by their interestingness.
3. A large collection of CCSK assertions for 386 subjects covering 3 domains (geography, religion, occupation) and several facets (food, drinks, clothing, traditions, rituals, behaviors).

Experimental evaluations show that the assertions in CANDLe are of significantly higher quality than those from prior works. An extrinsic use case demonstrates that our CCSK can improve performance of GPT-3 in question answering. Code and data can be accessed at https://candle.mpi-inf.mpg.de/.

2 RELATED WORK

Commonsense knowledge acquisition. There is a long tradition of CSK acquisition in AI (e.g., [10, 15, 19, 28, 34]). Earlier projects, e.g., Cyc [15] and ConceptNet [19], construct commonsense knowledge graphs (CSKGs) based on large-scale human annotations. Crowdsourcing CSKG construction has been revived in the ATOMIC project [13, 32]. CSK extraction from texts has been researched in WebChild [37], TupleKB [6], Quasimodo [30], ASER [46, 47], TransOMCS [45], GenericSK [3], and Ascent [21, 22]. Meanwhile, Ilievski et al. [14] consolidate CSK from 7 different resources into one integrated KG. Those projects, however, have their main focus on either concept-centered knowledge (e.g., Elephants have trunks), social interactions (e.g., X hates Y’s guts, as a result, X wants to yell at Y), or event-centered knowledge (e.g., X drinking coffee happens after X pouring the coffee into a mug) and do not cover much cultural knowledge. Our approach also starts from texts, but focuses on cultural commonsense knowledge (CCSK), with particular challenges in knowledge representation, assertion filtering and consolidation.

Cultural commonsense knowledge. A few works have focused specifically on CCSK. An early approach by Anacleto et al. [2] gathers CSK from users from different cultures, entered via the Open Mind Common Sense portal. However, the work is limited to a few eating habits (time for meals, what do people eat in each meal?, food for party/Christmas) in 3 countries (Brazil, Mexico, USA), and without published data. Acharya et al. [1] embark on a similar manual
These are the key labels for CCSK assertions. 16 geo-diverse commonsense concepts (e.g., traffic rules, dates for food assertions sometimes overlap with traditions. Note that the cultural facets need not be mutually exclusive, e.g., sentences with salient concepts.

3 major domains: geography, religion and occupation) and facets.

3 CCSK REPRESENTATION

Our representation of CCSK is based on the notions of subjects (from 3 major domains: geography, religion and occupation) and facets. These are the key labels for CCSK assertions, which are informative sentences with salient concepts marked up.

We assume two sets to be given:

- \( S \): A set of subjects (cultural groups) \( s_1, \ldots, s_n \) from a cultural domain, e.g., based on geo-locations (United States, China, Middle East, California), religious groups (Christians, Muslims, Buddhists) or occupations (taxi driver, professor, web developer);
- \( F \): A set \( F_1, \ldots, F_m \) of facets of culture, e.g., food, drinks, clothing, traditions, rituals, behaviors.

Note that the cultural facets need not be mutually exclusive, e.g., food assertions sometimes overlap with traditions.

Our objective is to collect a set of CCSK assertions for a given subject and facet. Existing commonsense resources store assertions in triple format (e.g., ConceptNet [36], Quasimodo [30]), semantic frames (Ascent [22]) or generic sentences (GenericsKB [3]). Although the traditional triple-based and frame-based data models are convenient for structured querying, and well suited for regular assertions like birth dates, citizenships, etc., they often falls short of capturing nuanced natural-language assertions, as essential for CSK. Moreover, recent advances in pre-trained LMs have made it easier to feed downstream tasks with less structured knowledge.

With CANDLE, we thus follow the approach of GenericsKB [3], and use natural-language sentences to represent assertions.

In principle, an assertion could comprise even several sentences. The longer the assertions are, however, the harder it is to discern their core. In this work, for higher precision and simplicity of computations, we only consider single sentences.

Definition 1 (Cultural commonsense knowledge assertion). Given a subject \( s \) and a facet \( F \), a CCSK assertion is a triple \(( s, F, \text{sent})\) where \( \text{sent} \) is a natural-language sentence about facet \( F \) of subject \( s \).

Since natural language often allows to express similar assertions in many different ways, and web harvesting naturally leads to discovering similar assertions multiple times, we employ clustering as an essential component in our approach.

A cluster \((\text{cls})\) of CCSK assertions for one subject and cultural facet contains assertions with similar meaning, and for presentation purposes, is summarized by a single summary sentence. Each cluster also comes with a score denoting its interestingness.

To further organize assertions, we also identify salient concepts, i.e., important terms inside assertions, that can be used for concept-centric browsing of assertion sets.

Several examples of CCSK assertions produced by CANDLE are shown in Fig. 2.

4 METHODOLOGY

We propose an end-to-end system, called CANDLE, to extract and organize CCSK assertions based on the proposed CCSK representation. Notably, our system does not require annotating new training data, but only leverages pre-trained models with judicious techniques to enhance the accuracy. The system takes in three inputs:

- an English text corpus (e.g., a large web crawl);
- a set of subjects (cultural groups);
- a set of facets of culture.

CANDLE consists of 6 modules (see Fig. 3). Throughout the system, step by step, we reduce a large input corpus (which could contain billions of documents, mostly noisy) into high-quality clusters of CCSK assertions for the given subjects and facets. Each cluster in the output is also accompanied by a representative sentence and an interestingness score. We next elaborate on each module.

4.1 Subject detection

We start the extraction by searching for sentences that contain mentions of the given subjects. These will be the candidate sentences used in the subsequent modules. To achieve high recall, we utilize generous approaches such as string matching and named entity
4.2 Generic assertion filtering

CSK aims at covering generic assertions, not episodic or personal experiences. For example, "Germans like their currywurst" is a generic assertion, but "I visited Germany to eat currywurst or this restaurant serves German currywurst" are not.

GenericsKB [3] is arguably the most popular work on automatically identifying generic sentences in texts and it uses a set of 27 hand-crafted lexico-syntactic rules. CANDLE adopts those rules in this module. However, for each domain and facet, we adaptively drop some of the rules if they would reject valuable assertions. More details on the adaptations can be found in Appx. B.

4.3 Cultural facet classification

To organize CCSK and filter out irrelevant assertions, we classify candidate sentences into several facets of culture. Traditional methods for this classification task would require a substantial amount of annotated data to train a supervised model. The costs of data annotation are often a critical bottleneck in large-scale settings. In CANDLE, we aim to minimize the degree of human supervision by leveraging pre-trained models for zero-shot classification.

A family of pre-trained models that is suitable for our setting is textual entailment (a.k.a. natural language inference - NLI); given two sentences, does one entail the other (or are they contradictory or unrelated)? Our approach to adopting such a model for cultural facet classification is inspired by the zero-shot inference method of Yin et al. [44]. Given a sentence sent and a facet F, we construct the NLI test as follows:

**Input:** Premise ← sent, Hypothesis ← “This text is about F”

**Output:** \[P(\text{sent } \in F) ← P(\text{Premise } \Rightarrow \text{Hypothesis})\]

The probability of Premise entailing Hypothesis will be taken as the probability of sent being labeled as F, denoted as \(P(\text{sent } \in F)\). For example, with sentence "German October festivals are a celebration of beer and fun", the candidate entailments will be "This text is about drinks", "... about food", "... about traditions", and so on. Multiple of these facets may yield high scores in these NLI tests.

To enhance precision, we introduce a set of counter-labels for topics that are completely outside the scope of CCSK, for example, politics or business. A sentence sent will be accepted as a good candidate for facet F if

\[
\begin{align*}
P(\text{sent } \in F) & \geq \rho_+, \text{ and } \\
\rho(\text{sent } \in \hat{F}) & \leq \rho_- \text{ for all counter-labels } \hat{F} 
\end{align*}
\]

where \(\rho_+\) and \(\rho_-\) are hyperparameters in the range \([0, 1]\), giving us the flexibility to tune for either precision or recall.

In our experiments, we use the BART model [16] finetuned on the MultiNLI dataset [41] for NLI tests. Our crowdsourcing evaluations show that the zero-shot classifiers with the enhanced techniques achieved high precision (see Appx. D.2).

4.4 Assertion clustering

The same assertion can be expressed in many ways in natural language. For example, "Fried rice is a popular Chinese dish" can also be written as "Fried rice is a famous dish from China or one of the most popular Chinese foods is fried rice." Clustering is used to group such assertions, which reduces redundancies, and allows to obtain frequency signals on assertions.

We leverage a state-of-the-art sentence embeddings method, SentenceBert [29], to compute vector representations for all assertions and use the Hierarchical Agglomerative Clustering (HAC) algorithm for clustering. Clustering is performed on assertions of each subject-facet pair.

**Cluster summarization.** Since each cluster can have from a few to hundreds of sentences, it is important to identify what those sentences convey, in a concise way.

[3] Model available at https://huggingface.co/facebook/bart-large-mnli
One way to compute a representative assertion for a cluster is to compute the centroid of the cluster, then take its closest assertion as the representative. Yet for natural-language data, this does not work particularly well.

In CANDLE, we therefore approach cluster summarization as a generative task, based on a state-of-the-art LM, GPT-3 [5] (see Appx. F for prompt template). Annotator-based evaluations show that GPT-generated representatives received significantly better scores than the base sentences in the clusters (see Sec. 6.1).

4.5 Concept extraction

While the cultural groups are regarded as subjects, concepts are akin to objects of the assertions. Identifying these concepts enables concept-focused browsing (e.g., browsing Japan assertions only about the Miso soup, etc.). We postulate that main concepts of an assertion cluster are terms shared by many members: We extract all n-grams (n = 1...3) of all assertions in a cluster (excluding subjects themselves, and stop words); and retain the ones that occur in more than 60% of the assertions. If both a phrase and its sub-phrase appear, we only keep the longer phrase in the final output. Noun-phrase concepts are normalized by singularization.

4.6 Cluster ranking and post-filtering

Ranking commonsense assertions is a crucial task. Unlike encyclopedic knowledge, which is normally either true or false, precision of CSK is usually not a binary concept, as it generalizes over many groups. With CANDLE, we aim to pull out the most interesting assertions for each subject, and avoid overly generic assertions such as "Chinese food is good or firefighters work hard", which are very common in the texts.

Extracting and clustering assertions from large corpora gives us an important signal of an assertion, its frequency. However, ranking based on frequency alone may lead to reporting bias. As we compile a CCSK collection at large scale, it also enables us to compute the distinctiveness of an assertion against others in the collection. The notion of these 2 metrics can be thought of as term frequency and inverse document frequency in the established TF-IDF technique for IR document ranking [35]. Besides frequency and distinctiveness, we score the interestingness of assertion clusters based on 2 other custom metrics: specificity (how many objects are mentioned in the assertion?) and domain relevance (how relevant is the assertion to the cultural facet?).

Frequency. For each subject-facet pair, we normalize cluster sizes into the range [0, 1], using min-max normalization.

Distinctiveness. We compute the IDF of a cluster $cls$ as follows:

$$IDF(cls) = \frac{\sum_{cls' \in CLS size(cls') \times \sigma(cls, cls')}}{\sum_{cls' \in CLS size(cls')} \times \sigma(cls, cls')}$$

(2)

where CLS is the set of all clusters for a given facet (e.g., food) and domain (e.g., geography>country), and

$$\sigma(cls, cls') = \begin{cases} 1 & \text{if } sim(cls, cls') \geq \theta \\ 0 & \text{otherwise} \end{cases}$$

(3)

Here, $sim(cls, cls')$ is the semantic similarity between the two clusters $cls$ and $cls'$, and $\theta$ is a predefined threshold. In CANDLE, to reduce computation, we approximate $sim(cls, cls')$ as the similarity between their summary sentences, which can be computed as the cosine similarity between their embedding vectors. When computing these embeddings, the subjects in the sentences are replaced with the same [MASK] tokens so that we only compare the expressed properties. Then, we normalize the logarithmic IDF values into the range $[0, 1]$ to get the distinctiveness scores of clusters.

Specificity. We compute the specificity of an assertion based on the fraction of nouns in it. Concretely, in CANDLE, the specificity of a cluster is computed as the specificity of its summary sentence.

Domain relevance. For each facet, we compute the domain relevance of a cluster by taking the average of the probability scores given to its members by the cultural facet classifier.

Combined score. The final score for cluster $cls$ is the average of the 4 feature scores. A higher score means higher interestingness.

Post-filtering. Lastly, to eliminate redundancies and noise, and further improve the final output quality, we employ a few hand-crafted rules:

- At most 500 clusters per subject-facet pair are retained, as further clusters mostly represent redundancies or noise.
- We remove clusters that have no concepts extracted, or that are based on too few distinct sentences ($\geq 2$ same sentences) or web source domains.
- We remove any cluster if either its summary sentence or many of its member sentences match a bad pattern. We compile a set of about 200 regular expression patterns, which were written by a knowledge engineer in one day. For e.g., we reject assertions that contain "the menu", "the restaurant" (likely advertisements for specific restaurants), or animal and plant breeds named after locations, such as "American bison", "German Shepherd", etc.

5 IMPLEMENTATION

Input corpus. In CANDLE, we use the broad web as knowledge source, because of its diversity and coverage, which are important for long-tail subjects. Besides the benefits, the most challenging problem when processing web contents is the tremendous amount of noise, offensive materials, incorrect information etc., hence, choosing a corpus that has been chiefly cleaned is beneficial. We choose the Colossal Clean Crawled Corpus (C4) [27] as our

| Facet       | Geography | Religions | Occupations |
|-------------|-----------|-----------|-------------|
| #A          | #C        | #A        | #C          | #A          | #C          |
| Food        | 12,981    | 9,750     | 9,837       | 511         |
| Drinks      | 5,923     | 3,079     | 3,321       | 227         |
| Clothing    | 1,237     | 1,695     | 4,367       | 278         |
| Rituals     | 8,007     | 74,651    | 22,581      | 1,253       |
| Traditions  | 13,068    | 68,202    | 2,798       |             |
| Behaviors   | -         | -         | -           | -           |
| Other       | -         | 60,483    | 2,292       | 159,239     | 5,461       |
| All         | 41,754    | 217,860   | 9,155       | 224,497     | 9,225       |
We perform the following evaluations:

6 EVALUATION

We provide statistics of the output in Table 1. In total, the resource contains 196 countries, 365 million English articles, each with text content and source URL. Before passing it to our system, we preprocessed all C4 documents using spaCy, which took 2 days on our cluster of 6K CPU cores.

**Subjects.** We collect CCSK for subjects from 3 cultural domains: geography (272 subjects), religions (14 subjects) and occupations (100 subjects). For geography, we split into 4 sub-domains: countries, continents, geopolitical regions (e.g., Middle East, Southeast Asia, etc.) and US states, which were collected from the GeoNames database\(^3\), which also provides alias names. We further enriched these aliases with demonyms from Wikipedia\(^4\).

**Facets of culture.** We consider 5 facets: food, drinks, clothing, rituals, and traditions (for geography/religion) or behaviors (for occupation), selected based on an article on facets of culture [23].

**Execution and result statistics.** After tuning the system’s hyperparameters on small withheld data (see Appx. C), we executed CANDLE on a cluster of 6K CPU cores (AMD EPYC 7702) and 40 GPUs (a mix of NVIDIA RTX 8000, Tesla A100 and A40 GPUs).

Regarding processing time, for the domain country (196 subjects), it took a total of 12 hours to complete the extraction, resulting in 8.4K clusters for the facet food (cf. Table 2). Occupations and religions took 8 and 6 hours each.

We provide statistics of the output in Table 1. In total, the resulting collection has 1.1M CCSK assertions (i.e., base sentences) which form 60K clusters for the given subjects and facets.

### Table 2: Processing time and output size of each step in CANDLE for the domain geography-country and facet food.

| #   | Step                          | Time   | Output/Data size          |
|-----|-------------------------------|--------|---------------------------|
| 1   | Subject detection             | 2 hours| 367M subject matches      |
|     |                               |        | 300M sentences (-93%)     |
| 2   | Generic assertion filtering   | 2 hours| 13M generic sentences     |
|     |                               |        | (-96%)                    |
| 3   | Cultural facet classification | 4 hours| 769K positive sentences   |
|     |                               |        | (-94%)                    |
| 4   | Assertion clustering          | 4 hours| 42K clusters (-93%)       |
| 5   | Concept extraction            | < 5 minutes| 12.4K concepts           |
| 6   | Cluster ranking and post-filtering | < 5 minutes| 8.8K clusters (-80%) |
|     | Total                         | ~ 12 hours | ~ 12 hours               |

In Appx. D, we also break down our CCSK collection into domains and facets, analyzing in details the assertion quality for each sub-collection.

6.1 Comparison with other resources

6.1.1 Evaluation metrics. Following previous works [7, 30], we analyze assertion quality along several complementary metrics, annotated by Amazon MTurk (AMT) crowdsourcing.

1. **Plausibility (PLA).** This dimension measures whether assertions are considered to be generally true, a CCSK-softened variant of correctness/precision.
2. **Commonality (COM).** This dimension measures whether annotators have heard of the assertion before, as a signal for whether assertions cover mainstream or fringe knowledge (akin to salience).
3. **Distinctiveness (DIS).** This dimension measures discriminative informativeness of assertions, i.e., whether the assertion differentiates the subject from others.

Each metric is evaluated on a 3-point Likert scale for negation (0), ambiguity (1) and affirmation (2). Distinctiveness (DIS) is only applicable if the answer to the plausibility (PLA) question is either 1 or 2. In case the annotators are not familiar with the assertion, we encourage them to perform a quick search on the web to find out the answers for the PLA and DIS questions. More details on the AMT setup can be found in Appx. E.

6.1.2 Compared resources. We compare CANDLE with 3 prominent CSK resources: Quasimodo [30], Acharya et al. [1], StereoKG [7]. The former covers broad domains including assertions for countries and religions, while the others focus on cultural knowledge.

Other popular resources such as ConceptNet [36], GenericsKB [5], Ascent/Ascent++ [21, 22], ATOMIC [32], ASER [47] and TransOMCS [45] do not have their focus on cultural knowledge and contain very little to zero assertions for geography or religion subjects, hence, they are not qualified for this comparison.

We evaluate 2 versions of CANDLE, one where each base assertion is retained independently (CANDLE-base-sent), the other containing only the cluster representatives (CANDLE-cluster-reps).

6.1.3 Setup. For comparability, all resources are compared on 100 random assertions of the same 5 country subjects covered in StereoKG [7] - United States, China, India, Germany and France. We note that among all compared resources, Acharya et al. [1] only contain two subjects (United States and India), so for that resource, we only sample from those. For StereoKG, we use their natural-language assertions. For Quasimodo and Acharya et al., we

---

\(^3\)[http://www.geonames.org/]
\(^4\)[https://en.wikipedia.org/wiki/Demonym]
verbalize their triples using crafted rules. Each assertion is evaluated by 3 MTurk annotators. Additionally, we ask if the annotator would consider the assertion as an inappropriate or offensive material. More details on the annotation task can be found in Appx. E.

6.1.4 Results. A summary of comparison with other resources is shown in Table 3.

Resource size and assertion length. CANDLE outperforms all other resources on the number of base sentences. When turning to clusters, our resource still has significantly more assertions than Acharya et al. (which was constructed manually at small scale) and StereokG (extracted from Reddit/Twitter questions). Quasimodo has comparable size with CANDLE-cluster-reps for the country and religion domains and has more for the occupation domain.

The OpenIE-based methods, Quasimodo and StereokG, produce the shortest assertion (32 and 37 characters on average, respectively). The manually-constructed KG (Acharya et al.) has the longest assertions (102 characters). CANDLE, having average assertion lengths (69 and 73), stands between those two approaches.

Assertion quality. In general, CANDLE-cluster-reps considerably outperforms all other baselines on 2 of the 3 metrics (plausibility and distinctiveness). Our resource only comes behind Acharya et al. on the commonality metric (1.15 and 1.22 respectively), which is expected because Acharya et al. only cover a few relations about common rituals (e.g., birthday, wedding) in 2 countries, USA and India, and their assertions are naturally known by many workers on Amazon MTurk, who are mostly from these 2 countries [31]. Importantly, the resource of Acharya et al. is based on crowdsourcing and only contains a small set of 225 assertions for a few rituals.

CANDLE-cluster-reps even outperform the manually-constructed KG (Acharya et al.) on the plausibility metric. This could be caused by an annotation task design that is geared towards abnormalities, or lack of annotation quality assurance.

CANDLE also has the highest scores on the distinctiveness metric, while most of the assertions in other resources were marked as not distinguishing by the annotators. Between the two versions of CANDLE, the cluster representatives consistently outperform the base sentences on all evaluated metrics. This indicates that still some of the raw sentences in the collection are noisy, on the other hand, the computed cluster representatives are more coherent and generally of better quality.

We also measured the offensiveness (OFF) of each resource, i.e., the percentage of assertions that were marked as inappropriate or offensive materials by at least one of the human-annotators. Quasimodo and StereokG, extracted from raw social media contents, have the highest number of assertions considered offensive (18% and 13%). Meanwhile, CANDLE’s judicious filters only miss a small fraction (1% of final assertions).

In summary, our CANDLE CSK collection has the highest quality by a large margin compared to other resources. Our resource provides assertions of high plausibility and distinctiveness. The clustering and cluster summarization also help to improve the presentation quality of the CSK.

6.2 Comparison with direct LM extraction

Knowledge extraction directly from pre-trained LMs is recently popular, e.g., the LAMA probe [25] or AutoTOMIC [40]. There are major pragmatic challenges to this approach, in particular, that assertions cannot be contextualized with truly observed surrounding sentences, and that errors cannot be traced back to specific sources. Nonetheless, it is intrinsically interesting to compare assertion quality between extractive and generative approaches. In this section, we compare CANDLE with assertions generated by the state-of-the-art LM, GPT-3 [5].

Generating knowledge with GPT-3. We query the largest GPT-3 model (davinci-002) with the following prompt template: “Please write 20 short sentences about notable <facet> in <subject>.” We run each prompt 10 times and set the randomness (temperature) to 0.7, so as to obtain a larger resource. We run the query for 5 facets and 210 subjects (196 countries and 14 religions), resulting in 188,061 unique sentences. Henceforth we call this dataset GPT-resource, and reuse it in the extrinsic use cases (Sec. 6.3).

Evaluation metrics and setup. For each resource, we sample 100 assertions for each facet (hence, 500 assertions in total) and perform human evaluation on the 3 metrics - commonality (COM), plausibility (PLA) and distinctiveness (DIS).

Results. The quality comparison between assertions of CANDLE and GPT-resource is shown in Table 4. While plausibility scores are the same, and CANDLE performs better in commonality, the
The Chinese wear new clothes for the New Year to symbolize new beginnings.

The bride usually wears red in a traditional Chinese wedding.

The costumes in Chinese opera are very colorful and important.

Chinese people also like to dress their children in very cute clothes.

The Chinese wear white at funerals because it is associated with mourning in Chinese culture.

In the winter, people in China wear coats and scarves to keep warm.

Shoes are also very important in Chinese culture.

In China, you will often see little girls wearing dresses and boys wearing shorts.

**Table 5: Example assertions of CANDLE and GPT-resource for subject: China, facet: clothing.**

| # | CANDLE                                                                 | GPT-resource                                                                 |
|---|------------------------------------------------------------------------|------------------------------------------------------------------------------|
| 1 | The bride usually wears red in a traditional Chinese wedding.        |                                                                 physicians also like to wear modern clothes such as jeans and t-shirts. |
| 2 | The Chinese wear white at funerals because it is associated with mourning in Chinese culture. |                                                                 Shoes are also very important in Chinese culture. |
| 3 | The Chinese wear new clothes for the New Year to symbolize new beginnings. |                                                                 Chinese people also like to dress their children in very cute clothes. |
| 4 | The costumes in Chinese opera are very colorful and important.        |                                                                 In China, you will often see little girls wearing dresses and boys wearing shorts. |
| 5 | In ancient China, only the emperor was allowed to wear the color yellow. |                                                                 In the winter, people in China wear coats and scarves to keep warm. |

| Facet | #Questions | Precision (%) | No cont. | GPT cont. | CANDLE cont. |
|-------|------------|---------------|----------|-----------|-------------|
| Food/Drinks | 88 | 92.05 | 94.32 | 93.18 |
| Behaviors | 125 | 60.80 | 57.60 | 63.20 |
| Rituals | 135 | 87.41 | 85.93 | 92.59 |
| Traditions | 152 | 72.37 | 69.74 | 79.61 |
| All | 500 | 77.00 | 75.40 | 81.40 |

6.3 Extrinsic evaluation

**Table 6: Results of QA using context-augmented LMs.**

| Facet | #Questions | Precision (%) | No cont. | GPT cont. | CANDLE cont. |
|-------|------------|---------------|----------|-----------|-------------|
| Food | 63.0 | 30.0 | 44.0 | 70.0 | 84.0 | 58.2 |
| Drinks | 74.0 | 62.0 | 76.0 | 80.0 | 75.4 |

As input, we select a random set of 100 countries, and take assertions from either CANDLE or GPT-resource. The game has 5 rounds, each is associated with a facet of culture. In each round, for each country, we draw the top-5 assertions from each resource (sorted by interestingness in CANDLE or by frequency in GPT-resource). All mentions of the countries in the input sentences are replaced with [ . . . ], before being revealed to the player.

This is a game that requires a player that possesses a wide range of knowledge across many cultures. Instead of human players, we choose GPT-3 as our player, which has been shown to be excellent at many QA tasks [5] (prompt settings are presented in Appx. F).

We measure the precision of the answers and present the results in Table 7. It can be seen that the player got significantly more correct answers when given assertions from CANDLE than from GPT-resource (i.e., assertions written by the player itself!). This confirms that assertions in CANDLE are more informative.

7 CONCLUSION

We presented CANDLE—an end-to-end methodology for automatically collecting cultural commonsense knowledge (CCSK) from broad web contents at scale. We executed CANDLE on several cultural subjects and facets of culture and produce CCSK of high quality. Our experiments showed the superiority of the resulting CCSK collection over existing resources, which have limited coverage for this kind of knowledge, and also over methods based on prompting LMs. Our work expands CSKG construction into a domain that has been largely ignored so far. Our data and code are accessible at https://candle.mpi-inf.mpg.de/.

Ethics statement

No personal data was processed and hence no IRB review was conducted. It is in the nature of this research, however, that some outputs reflect prejudices or are even offensive. We have implemented multiple filtering steps to mitigate this, and significantly reduced the percentage of offensive assertions, compared with prior work. Nonetheless, CANDLE represents a research prototype, and outputs should not be used in downstream tasks without further thorough review.
A SYSTEM SUMMARY

We provide an overview of techniques and models applied in each module of the CANDLE system in Table 8.

Table 8: List of models and techniques used in each module of CANDLE.

| # | Module             | Used techniques/models                                      |
|---|--------------------|-------------------------------------------------------------|
| 1 | Subject detection  | String matching, NER                                        |
| 2 | Generic assertion  | Hand-crafted lexico-syntactic rules                         |
| 3 | Cultural facet     | Huggingface’s bar-t-large-mr@1 model for the textual entailment task |
| 4 | Assertion clustering | SentenceBert for clustering, GPT-3 for cluster summarization |
| 5 | Concept extraction | Common n-gram extraction                                   |
| 6 | Cluster ranking + post-filtering | Ad-hoc ranking features, rule-based filtering |

B GENERIC FILTERING RULES

GenericsKB [3] was built by using a set of 27 hand-crafted lexico-syntactic rules to extract high-quality generic sentences from different text corpora (the ARC corpus, SimpleWikipedia and the Waterloo crawl of education websites). For example, the lexical rules look for sentences with short length, starting with a capitalized character, having no bad first words (e.g., determiners), ending with a period, having no URL-like snippets, etc. The syntactic rules only accept a sentence if its root is a verb and not the first word, and if there is a noun before the root verb, etc.

CANDLE adopts the GenericsKB rules. However, as GenericsKB only deals with general concepts (e.g., “tree”, “bird”, etc.), some of the rules are not applicable for the cultural subjects that can be named entities. Hence, depending on the subjects and facets, we adaptively modify the rules (by dropping some of them) so that we will not miss out valuable assertions. For instance, for geography, the has-no-determiners-as-first-word rule will filter out valuable assertions such as “The Chinese use chopsticks to eat their food” or “The currywurst is a traditional German fast food dish, and it must be dropped. In another situation, when exploring the “traditions” facet, the remove-past-tense-verb-roots rule would be too aggressive as it rejects assertions about past traditions. The rule that rejects sentences with PERSON entities can be used for the geography and occupation subjects, but must not be used for religions, because it will filter out sentences about Buddha or Jesus Christ. Full details are in the published code base5.

C HYPERPARAMETER SETTINGS

Based on tuning on small withheld data, we select the following values for hyperparameters and run CANDLE on the C4 dataset with these settings. For cultural facet classification (cf. Sec. 4.3 and Eq. 1), we fix $\rho_5 = 0.5$ and $\rho_6 = 0.3$. For assertion clustering (cf. Sec. 4.4), we use the SentenceBert model all-MiniLM-L6-v2 for computing sentence embeddings. For the HAC algorithm, we measure point-wise Euclidean distance of the normalized embeddings. Then, we use the Ward’s linkage [38], with the maximal distance threshold set to 1.5. In the few cases where input sets are larger, we truncate them at 50K sentences per subject-facet pair, since larger inputs only contain further redundancies, that are not worth the cubic effort of clustering. This concerns only 15 out of 386 subjects. For cluster summarization, we consider the 500 most populated clusters for each subject-facet pair with a minimum size of 3 sentences. For cluster ranking (cf. Sec. 4.6), we fix $\theta$ in Eq. 3 to 0.8.

D INTRINSIC EVALUATION

We break down the CANDLE CCKS collection into domains and facets and evaluate the assertion quality for each of these sub-collections and get more insights into the produced data.

D.1 Per-domain quality

CANDLE contains 3 cultural domains - geography, religion and occupation. For each domain, we sample 100 assertions and perform crowdsourcing evaluation with the 3 metrics - PLA, COM and DIS (cf. SubSec. 6.1.1). We present the evaluation results in Table 9. Besides the raw scores (0, 1, 2), we also binarize and denote them as acceptances, i.e., a score greater than zero means “accept”.

CANDLE achieves a high plausibility (PLA) score of 1.54 on average. Performance on this metric is relatively consistent throughout all domains. Meanwhile, the commonality (COM) metric is highest for the occupation domain and lowest for geography domain. More than 80% of plausible assertions are annotated as distinctive (DIS). Religion and occupation assertions perform significantly better than geography’s on this metric. That could be caused by several assertions for geography subjects being correct but too generic (e.g., “Japanese food is enjoyed by many people”, or “German beer is good”). On the other hand, religions and occupations are more distinguishing from one another, while countries or geo-regions usually have cultural overlaps.

D.2 Per-facet quality

We select the assertions for the domain country, and for each facet (food, drinks, clothing, traditions, rituals) we sample 100 assertions for crowdsourcing evaluation. Besides commonality (COM), plausibility (PLA) and distinctiveness (DIS), here we introduce one more evaluation metric, domain relevance (DOM), to measure if an assertion talks about the cultural facet of interest. Only when the DOM score is greater than zero, the other metrics will be evaluated. We present the evaluation results in Table 10.

---

5https://github.com/cultural-csk/candle/blob/main/candle/pipeline/component_generic_sentence_filter.py
Table 10: Quality of CANDLE assertions for each facet and the domain geography>country.

| Facet       | Quality [0..2] |
|-------------|---------------|
|             | DOM | PLA | COM | DIS |
| Food        | 1.42 | 1.23 | 0.94 | 0.97 |
| Drinks      | 1.51 | 1.40 | 1.14 | 1.19 |
| Clothing    | 1.49 | 1.30 | 1.04 | 1.07 |
| Rituals     | 1.45 | 1.27 | 1.06 | 1.20 |
| Traditions  | 1.42 | 1.27 | 1.02 | 1.11 |
| Average     | 1.46 | 1.29 | 1.04 | 1.11 |

It can be seen that CANDLE maintains good quality on all evaluation metrics. Notably, scores for the DOM metric are consistently high for all facets, suggesting that the enhanced techniques for zero-shot classification work well on our data. Interestingly, the facet drinks outperforms all other facets on 3 of the 4 metrics (DOM, PLA and COM), especially for PLA, its score is significantly higher than others. Assertions for drinks and rituals are also more distinctive than for other facets.

E DETAILS OF ANNOTATION TASK FOR ASSERTION EVALUATION

The evaluations of assertion quality (Tables 3, 4, 9 and 10) are conducted on Amazon MTurk (AMT). We present CCSK assertions to annotators in the form of natural-language sentences (triples from Quasimodo [30] and Acharya et al. [1] were verbalized using crafted rules). We evaluate each assertion along 3-4 dimensions on a 3-point Liker scale - negation (0), ambiguity (1) and affirmation (2). Each AMT task consists of 5 assertions evaluated by 3 different annotators. Workers are compensated $0.50 per task. We select Master workers with lifetime’s acceptance rate more than 99%. We obtain fair inter-annotator agreements given by Fleiss’ kappa [9]: 25.0 for DOM, 25.7 for PLA and 25.4 for DIS. This number for COM (13.4) is lower than others because it is an objective question (has the annotator heard of the assertion?).

F GPT-3 PROMPTING

In this work, we use GPT-3 for cluster summarization (Sec. 4.4), generating CCSK for GPT-resource (Sec. 6.2), context-augmented QA and “guess the country” game (Sec. 6.3). The prompt templates and settings used for these tasks are presented below.

Cluster summarization. We query the curie-001 model, with zero temperature and maximum length of 50 tokens. We only take the first generated sentence as output. An example prompt is presented in Fig. 4.

Generating CCSK for GPT-resource. We use the davinci-002 model and set temperature to 0.7 and maximum length to 512 tokens. For each facet and subject, we run the following prompt template for 10 times: Please write 20 short sentences about notable <facet> in <subject>. We query for 5 facets (food culture, drinking culture, clothing habits, rituals, traditions), and 210 subjects (196 countries and 14 religions). Table 5 shows some example generations for the subject China and the facet “clothing habits”.

It can be seen that CANDLE maintains good quality on all evaluation metrics. Notably, scores for the DOM metric are consistently high for all facets, suggesting that the enhanced techniques for zero-shot classification work well on our data. Interestingly, the facet drinks outperforms all other facets on 3 of the 4 metrics (DOM, PLA and COM), especially for PLA, its score is significantly higher than others. Assertions for drinks and rituals are also more distinctive than for other facets.

Figure 4: A screenshot of GPT-3 output for cluster summarization.

Figure 5: Screenshots of GPT-3 output in the QA task, without and with CCSK.

Context-augmented QA. We query the davinci-002 model with zero temperature and maximum length of 16 tokens. Answers are then manually mapped to the respective options. Example prompts are shown in Fig. 5.

“Guess the country” game. We use the davinci-002 model, with temperature=0 and a max_length=8. Answers given by GPT-3 are checked manually for their correctness. Example prompts can be seen in Fig. 6.